Registration of Synthetic Aperture Imagery using Feature Matching

Victor T. Wang
B.E. (Hons I)

A thesis presented for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering
at the
University of Canterbury,
Christchurch, New Zealand.

February 2018

Abstract

An important application in sonar research is change detection, which is reliant on accurate alignment of repeat-pass images. Although less accurate than correlation-based methods, feature-based methods are gaining popularity in radar and sonar imaging due to their relative computational efficiency. This thesis explores the feasibility of image registration of speckled imagery using feature matching.

As a proof of concept, a feature-based registration method is proposed for synthetic aperture sonar (SAS) based on an ideal sonar track. The registration pipeline uses the Scale Invariant Feature Transform (SIFT) and RANSAC estimation, a popular combination in computer vision. Using a novel track registration approach, feature correspondences are used to estimate a set of track registration parameters from which an image registration can be computed. This method produced an accurate alignment to within 0.03 pixels for a simulated repeat-pass scene with a 0.2 m baseline.

Desirable aspects of feature matching performance include a sufficient density of detected features, a high ratio of inlier matches, and accurate feature localisation. Since the performance of feature matching is known to be situational, simulated data was used to produce a large number of images from which general trends could be observed. The effect of sinc-interpolated sub-pixel shifts on feature matching performance was measured, with non-oversampled speckle images yielding an inlier ratio of less than 1 % in the worst case. Features were significantly more robust for oversampled images, with the expected worst-case inlier ratio being around 45 % for two-times oversampled images and around 77 % for four-times oversampled images. Overall, oversampled images were shown to provide better feature repeatability, increased density of features, and lower correspondence localisation errors compared to images without oversampling.

SIFT and SURF (Speeded Up Robust Features) were evaluated for performance in relation to speckle decorrelation and scene content. Feature matching was shown to be problematic for bland scenes with coherence below 0.9. Feature matching on non-bland scenes yielded more features, increased feature repeatability, and improved localisation accuracy. A model is proposed to capture the trend between feature repeatability and scene coherence for any given feature matching pipeline. Although the feasibility of feature matching in a given scenario cannot easily be predicted, a low number of matches and moderate localisation errors are both likely to have unfavourable implications on success and/or registration accuracy.

Deputy Vice-Chancellor's Office
Postgraduate Office

UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Co-Authorship Form

This form is to accompany the submission of any thesis that contains research reported in co-authored work that has been published, accepted for publication, or submitted for publication. A copy of this form should be included for each co-authored work that is included in the thesis. Completed forms should be included at the front (after the thesis abstract) of each copy of the thesis submitted for examination and library deposit.

Publication 1
Wang, V. and Hayes, M. P. (2014). Image registration of simulated synthetic aperture sonar images using SIFT. In *Image and Vision Computing New Zealand*, pages 31–36. ACM.
Early sections of Chapter 5 present material from this publication.
75% of the research/analysis and 90% of the writing was contributed by the candidate.

Publication 2
Wang, V. and Hayes, M. P. (2017b). Synthetic aperture sonar track registration using SIFT image correspondences. *IEEE Journal of Oceanic Engineering*, 42(4):901–913.
Chapter 5 presents material from this publication.
80% of the research/analysis and 95% of the writing was contributed by the candidate.

Publication 3
Wang, V. and Hayes, M. (2017a). SIFT localisation accuracy on interpolated speckle images. In *Image and Vision Computing New Zealand*. IEEE.
Chapter 6 presents material from this publication.
90% of the research/analysis and 95% of the writing was contributed by the candidate.

Publication 4
Wang, V. and Hayes, M. P. (2016b). Analysis of feature matching performance on correlated speckle image pairs. In *OCEANS*. MTS/IEEE.
Early sections of Chapter 7 present material from this publication.
80% of the research/analysis and 95% of the writing was contributed by the candidate.

Publication 5
Wang, V. and Hayes, M. (2016a). Modelling of feature matching performance on correlated speckle images. In *Image and Vision Computing New Zealand*. IEEE.
Chapter 7 presents material from this publication.
90% of the research/analysis and 95% of the writing was contributed by the candidate.

Certification by Co-authors:
If there is more than one co-author then a single co-author can sign on behalf of all.
The undersigned certifys that:

- The above statement correctly reflects the nature and extent of the PhD candidate's contribution to this co-authored work.
- In cases where the candidate was the lead author of the co-authored work he or she wrote the text.

Name: *Michael Hayes* Signature: *Michael Hayes* Date: *28 February 2018*

Acknowledgements

Firstly, I would like to thank my supervisor Dr. Michael Hayes for his guidance and support. His patience in explaining concepts and the general liveliness of his character are among his numerous admirable qualities that have helped me reach this point of being able to close this chapter of my life. Thanks also to my co-supervisor Dr. Rick Millane for his involvement. I thank the US Office of Naval Research for supporting this research through grant N62909-11-1-7037 (change detection using multiple pass InSAS). An acknowledgement is given to Naval Surface Warfare Center: Panama City Division for providing simulated SAS data.

A thank you to my family for enduring my prolonged partial absence, especially to my parents for supporting me in my journey of self-discovery and to my brother for taking up the slack where I have been unable. I am also grateful for receiving encouragement from friends and their oft-misplaced confidence in me; one time or another, both have helped me to bounce back after faltering.

Preface

This research began with the simple idea of demonstrating whether image registration based on feature matching could produce accurate results while significantly reducing processing time compared to correlation-based approaches. While results with a high quality simulated SAS dataset appeared to indicate feasibility, there were many questions left unanswered regarding reliability, repeatability, simplifying assumptions, and conditions for success. Our attempts to acquire more high quality SAS imagery (both simulated and real) were ultimately unfruitful, and thus we were unable to demonstrate our proposed feature-based track registration approach more convincingly.

However, from the results obtained, we glimpsed the element of unpredictability of the performance of feature matching, which could not be assessed or characterised without a large number of independent trials. By no coincidence, there were notable gaps in the literature concerning such measurements that are surely of some consequence, whether with sonar, radar, or optical image registration. The gap, the lack of statistics and predicted performance of feature matching, is due to the lack of ground-truth data in real applications. Our research turned towards addressing this topic and finding answers by means of generating random data sets with known ground truth. Although the assumptions made have been fairly idealistic, it is quite natural to first identify ideal performance bounds before extrapolating to more practical conditions that are highly dependent on each unique scenario.

Ultimately, this thesis encapsulates the earliest work (to our knowledge) that focuses primarily on feature-based registration of SAS imagery and the behavior of feature matching on speckled imagery in general.

Abbreviations

ADAC	automatic detection and classification
AUV	autonomous underwater vehicle
CCD	coherent change detection
CLT	central limit theorem
CRLB	Cramér-Rao lower bound
CT	computed tomography
DoG	difference of Gaussian
DoH	determinant of the Hessian
DVL	Doppler velocity log
FFT	fast Fourier transform
GPS	global positioning system
IMU	inertial measurement unit
InSAS	interferometric synthetic aperture sonar
laser	light amplification by stimulated emission of radiation
LoG	Laplacian of Gaussian
ML	maximum likelihood
OpenCV	open source computer vision
PDF	probability density function
radar	radio detection and ranging
RANSAC	random sample consensus
ROC	receiver operating curve

SAR	synthetic aperture radar
SAS	synthetic aperture sonar
SEM	standard error of the mean
SIFT	scale invariant feature transform
SNR	signal to noise ratio
sonar	sound navigation and ranging
SURF	speeded up robust features
WSS	wide-sense stationary

Contents

Chapter 1

Introduction

Change detection is an important application in sonar research, with ongoing developments leading towards automatic capabilities such as detection of buried mines from images obtained via periodic surveys of the seafloor. The most powerful form of change detection is image-based coherent change detection, which can reveal subtle and even visually imperceptible differences in a scene over time. However, the feasibility of such methods is completely reliant on accurate alignment of the repeat-pass images. Synthetic aperture sonar images are prone to distortion and artefacts due to imperfect navigation and other practical limitations, such that achieving artefact-free images and aligning them perfectly remains an ambituous endeavour in general. This thesis explores the topic of registering speckled images such as sonar and radar imagery from a computer vision approach using local image features. While feature-based methods are widely used for image registration of optical images, these techniques are yet to gain popularity with speckled imagery, especially in sonar, where its capabilities and performance implications are unclear. Although feature-based methods are known to be inferior in alignment accuracy compared to traditional correlation-based methods, they potentially offer greatly reduced computation times; clarification of the performance implications could lead to the development of hybrid algorithms or use in real-time applications.

This introductory chapter provides an overview of sonar, beginning with a brief history of echo detection and sonar in Sections 1.1 and 1.2. Section 1.3 explains the basic workings of a conventional side-scan sonar system. Section 1.4 describes how synthetic aperture processing can be used to achieve a higher resolution than with conventional sonar. Section 1.5 defines the concepts of range resolution and pulse compression, and the significance of sampling constraints in relation to aliasing is described in Section 1.6. Sections 1.7 and 1.8 discuss the formation of synthetic aperture sonar images and the need for motion compensation to improve image quality. The importance of change detection and image registration is explained in Section 1.9. An outline of the remainder of the thesis is given in Section 1.10. Finally, publications submitted as a part of this thesis work are listed in Section 1.11.

1.1 Echo detection

Much of the earth consists of aquatic environments, with around 70% of the surface being covered in water [Chang 2006]. Current knowledge and understanding of what takes place below the surface remains fairly limited. One assertion is that more is known about the moon than the deep sea, with over 80% of the ocean floor yet to been mapped using sonar. The most accurate global mapping of the ocean seafloor has a resolution of around 5 km; this was performed using radar to accurately measure the sea surface and infer the topography of the ocean floor [Sandwell et al. 2014]. Other imaging options are limited. Light attenuation in seawater is significant, where common visibilities range from tens of metres to less than a metre in turbid environments (typical of harbours) [Byrne et al. 2017]. With optical imaging being problematic, the remaining feasible option for high resolution charting of underwater scenes is sonar imaging. This is possible due to the relatively low attenuation of acoustic signals (sound) in water, especially with low frequency signals (below 1 kHz).

Echo detection is the use of sound to detect and locate objects. It has been known since Aristotle (384–322 B.C.E.) that sound can be heard in water as well as air. There is also some evidence of Phoenician fisherman (circa 500 B.C.E.) using echoes of ringing bells and other objects to detect nearby headlands [Kaharl 1999]. In 1490, Leonardo da Vinci described the use of a tube inserted in the water to detect noises from other ships [Fahy and Walker 1998].

Echolocation (a form of active sonar) is where an animal emits sounds into the environment, identifying objects and sensing distances based on the echoes they hear for the purposes of navigation and foraging/hunting. Echolocating animals use the differences in echoes heard from both ears, namely loudness and time delay, to perceive distance and direction. Examples of animals that use echolocation include microbats, toothed whales (such as dolphins, porpoises, and killer whales), shrews, and swiftlets. There are also reports of blind humans using sound to locate objects dating back as early as 1749 [Kolarik et al. 2014].

The first patent for an underwater device for echo-ranging was filed shortly after the sinking of the Titanic in 1911 [Urick 1975]. Further research in echo detection was driven by the need to detect enemy submarines during World War I, and led to the development of a wide range of techniques such as steerable hydrophone arrays.

1.2 Sonar

The acronyms SONAR (SOund Navigation And Ranging) and RADAR (RAdio Detection and Ranging) came into use during World War II, during which many new systems were developed including detection of naval mines using sonar and both active

and passive acoustic homing torpedoes [Urick 1975]. Post World War II, sonar came into use for civilian applications such as seafloor imaging and fish finding [Tucker 1966].

Active sonar systems typically use a projector to generate sound waves, where each distinct pulse of sound is called a *ping*. One or more hydrophones are used to measure the incoming reflections from these pings. The strength of an echo coming off a point in the scene depends on the texture and size/shape of the surface, the incoming angle, and the distance. Different sediments have different densities and reflectances. Specular reflections from mirror-like surfaces are strong only at specific viewing angles, whereas points and corners of objects scatter over a wide range of angles. Occluded regions (corresponding to shadows) have weak responses.

Seafloor imaging using side-scan sonar is now used for both military and civilian applications such as mine detection, shipwreck hunting, and pipe surveying. With side-scan geometry, the sound waves are propagated perpendicular to the direction of travel of the vessel. Sonar systems may be mounted on the vessel (typically a ship or submarine) or towed behind it (especially for shallow waters). This thesis focuses on the context of synthetic aperture sonar (SAS) imagery collected from side-scan sonar, however, many of the observations and findings relate to speckled imagery in general, including other sonar imaging modes, radar, and medical ultrasonography.

1.3 Conventional side-scan sonar

Conventional side-scan sonar (or narrow-beam sonar) operates by transmitting a narrow beam of acoustic energy to "illuminate" the seafloor scene. The echo returns are recorded for each individual pulse, with each sample corresponding to one strip of the sonar image being formed (see Figure 1.1). The travel path of the sonar is called the *sonar track*. The *along-track* direction is in the direction of travel (parallel to the y-axis), and the *across-track* direction is in the perpendicular horizontal direction (parallel to the x-axis). The *range* or *slant range* refers to the distance from the sonar to a point target, which is related to the across-track distance and the sonar altitude by the Pythagorean theorem. Narrower acoustic beams give better along-track resolution, as the along-track resolution is proportional to the beamwidth. The relationship can be approximated as [Caprais and Guyonic 1997]

$$\delta_y \approx \frac{r\lambda}{D}, \tag{1.1}$$

where r is the range of the target, λ is the wavelength of the transmitted beam, and D is the along-track width of the projector/hydrophone aperture. The along-track resolution is range-dependent; resolution is poorer for targets in the scene that are further away. The resolution can be improved by increasing the frequency of the acoustic waves (with the trade-off of increased signal attenuation at higher frequencies) or by using

longer apertures (subject to cost/manufacturing limitations).

High-resolution conventional side-scan systems use high frequencies but typically operate at short ranges due to signal attenuation. Narrow-beam sonar can be susceptible to parts of a scene being missed completely due to minor path variations. This problem, as well as different frequencies being capable of distinguishing different details, has led to the development of dual-frequency sonars such as described in [Sammelmann et al. 1997].

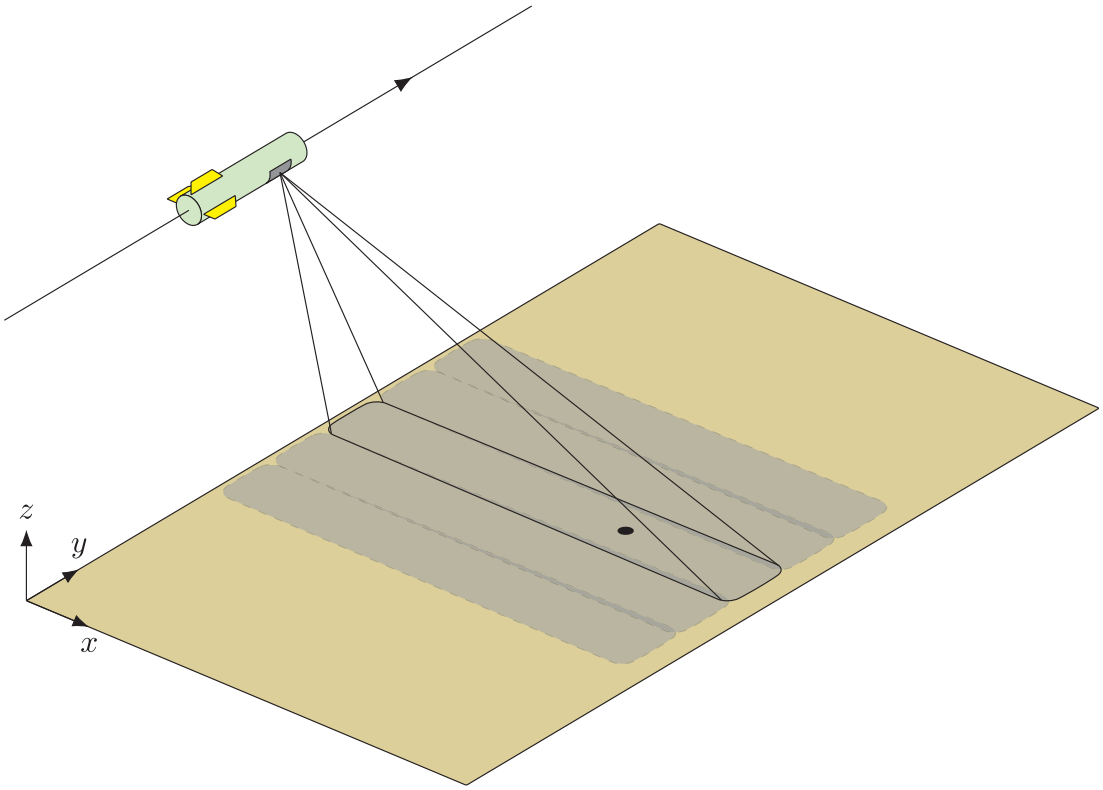


Figure 1.1: Imaging geometry of a conventional side-scan sonar, where each narrow-beam ping illuminates a strip of the seafloor.
Reprinted, with permission, from [Bonnett 2017].

1.4 Synthetic aperture sonar

Whereas conventional sonar uses a narrow acoustic beam, synthetic aperture sonar (SAS) uses a wide beam and combines the echoes from multiple pings using coherent processing to obtain a high resolution image. This method of processing effectively creates a synthetic (as opposed to real) aperture that is longer than the physical aperture of the transmitter, achieving a range-independent along-track resolution by varying the synthetic aperture length with range. For a SAS system, the along-track resolution is

given by

$$\delta_y = \frac{r\lambda}{2L}, \tag{1.2}$$

where L is the length of the synthetic aperture. A target that is further away is illuminated by more pings; thus, the effective aperture is proportional to the range of the target, as in

$$L = \frac{r}{\lambda D}. \tag{1.3}$$

Note that D is the width of either the projector or hydrophone aperture, whichever is greater. Combining with (1.2) gives an along-track resolution of

$$\delta_y = \frac{D}{2}. \tag{1.4}$$

A shorter physical aperture corresponds to a wider beam and a longer synthetic aperture, improving the along-track resolution. The resolution is also independent of frequency, allowing low-frequency systems to also generate high-resolution images. Although capable of attaining ten times the resolution of conventional sonar [Gendron et al. 2009][Sternlicht and Pesaturo 2004][Wille 2005], SAS does have some notable disadvantages. Echoes must have phase coherency for the length of the synthetic aperture, otherwise the resulting SAS image is severely degraded. Echo coherency implies adequate sampling as well as strict motion requirements that cannot always be guaranteed. SAS also requires heavy computation, with images being more complex to process, making it challenging to use for real-time applications.

Synthetic aperture imaging in sonar and radar has two modes: spotlight and stripmap. In spotlight mode, the beam is steered towards a small target area as the sonar system passes, whereas in stripmap mode the beam is always perpendicular to the direction of travel and does not change. SAS systems almost exclusively use stripmap imaging due to the navigational requirements for spotlight imaging (which are easily met for satellite SAR systems).

1.5 Range resolution and pulse compression

The range resolution is the minimum spacing between point targets, in terms of range, such that the two targets can be distinguished rather than detected as a single larger target. The range resolution of an echo detection system is proportional to the duration of the sound pulse used, and thus a shorter pulse achieves better range resolution. However, shorter pulses require more instantaneous power to generate enough energy, otherwise the system suffers from a low signal to noise ratio (SNR). Furthermore, there is a inherent limit on the amplitude that can be generated in water due to non-linear

effects such as cavitation [Urick 1975].

Pulse compression is a technique that allows good range resolution to be achieved while using longer pulses. The transmitted pulse uses a wide-band signal and the echo is correlated with the signal (or equivalently, match filtering is performed). The autocorrelation of a signal with bandwidth B has a time duration inversely proportional to the bandwidth, where the range resolution of a pulse-compressed echo is

$$\delta_r = \frac{c}{2B}. \tag{1.5}$$

Therefore, transmitting a longer pulse-compressed signal allows more energy to be transmitted (which increases the system SNR) while improving the range resolution. A common signal in sonar and radar used with pulse compression is the *linear chirp*, which is a sinusoidal wave with instantaneous frequency varying linearly with time. The factor of increase in system SNR is proportional to the product of duration and bandwidth of the chirp [Cook and Bernfeld 1967].

1.6 Sampling constraints

A synthetic aperture is a spatially sampled array and should be sampled at above the Nyquist rate, otherwise false targets appear in the reconstructed image due to aliasing. The along-track sample spacing is

$$\Delta y = \frac{v}{f_p}, \tag{1.6}$$

where v is the speed of the sonar and f_p is the temporal ping rate.

Based on the along-track resolution, it would seem that a sample spacing of

$$\Delta y \leq \frac{D}{2} \tag{1.7}$$

would be adequate—and this is a canonical sampling constraint in the literature [Douglas and Lee 1992][Tomiyasu 1978]. However, a rectangular aperture has an infinite spatial frequency extent, and thus aliasing is unavoidable. Furthermore, $D/2$ sampling has been shown to result in aliasing [Rolt and Schmidt 1992] even within the main lobe of the beam pattern. (The reader is referred to [Hawkins 1996] for a background on array theory.) $D/4$ sampling has been proposed as a more faithful sampling constraint, limiting aliasing to the side lobes of the beam pattern [Hawkins 1996].

The sampling rate imposes limitations on the sonar travel speed in combination with the ping rate. A slow speed is undesirable due to greater susceptibility to the effects of inevitable path deviations, and thus SAS systems usually use an array of hydrophones with along-track spacing between them [Douglas and Lee 1993]. With multiple receivers arranged this way, the sampling constraint is relaxed by a factor of

N, the number of receivers. This comes at the cost of increased complexity in synthetic aperture reconstruction [Callow 2003].

1.7 Image reconstruction

Image reconstruction is the process of forming the synthetic aperture to produce an output image using the collected SAS data. Unlike conventional sonar, which is non-coherent and produces images with real values, synthetic aperture processing is coherent, i.e., it retains the phase information of echoes to form complex images. There are multiple algorithms for SAS image reconstruction, including time-domain correlation [Nielsen 1991], backprojection, and the wavenumber algorithm. The reader is referred to [da Silva 2009] or [Callow 2003] for a more detailed background on reconstruction algorithms.

Backprojection is used in medical imaging to perform synthetic aperture reconstruction on computed tomography (CT) data and is also used for SAR and SAS image reconstruction. The algorithm works by *back-projecting* the echoes for each ping onto a spherical arc for all the contributing points in the image for a given echo. This is performed for every ping, and the accumulated result is the reconstructed image. The backprojection algorithm is faster than correlation and also offers the option of reconstructing data from an arbitrary track, which is not possible when using wavenumber reconstruction.

The resolution of a sonar system is the smallest target that the sonar can clearly resolve and is characterised by the along-track resolution and the across-track resolution (closely related to the range resolution). This minimum target area is called a *resolution cell*. The system resolution is dependent on the physical properties and specification of the sonar system, as well as the operating environment. In sonar, it is typical to have the image resolution match the system resolution such that each pixel corresponds to a resolution cell.

1.8 Motion compensation

Retaining phase coherency over the length of a synthetic aperture imposes a strict theoretical positioning accuracy along an ideal sonar path that is virtually impossible to achieve with free-towed systems and autonomous underwater vehicles (AUVs); even the most accurate rail-based systems operating under benign conditions do not necessarily achieve diffraction-limited images. Navigation errors are unpreventable due to the uncontrollable nature of the ocean environment and can cause severe degradation in image quality, including artefacts such as blurring and distortion. Of the six types of motion deviations from an ideal sonar path Figure 1.2, along-track motion (sway) and yaw (in the case of multiple-receiver systems) are the most problematic forms

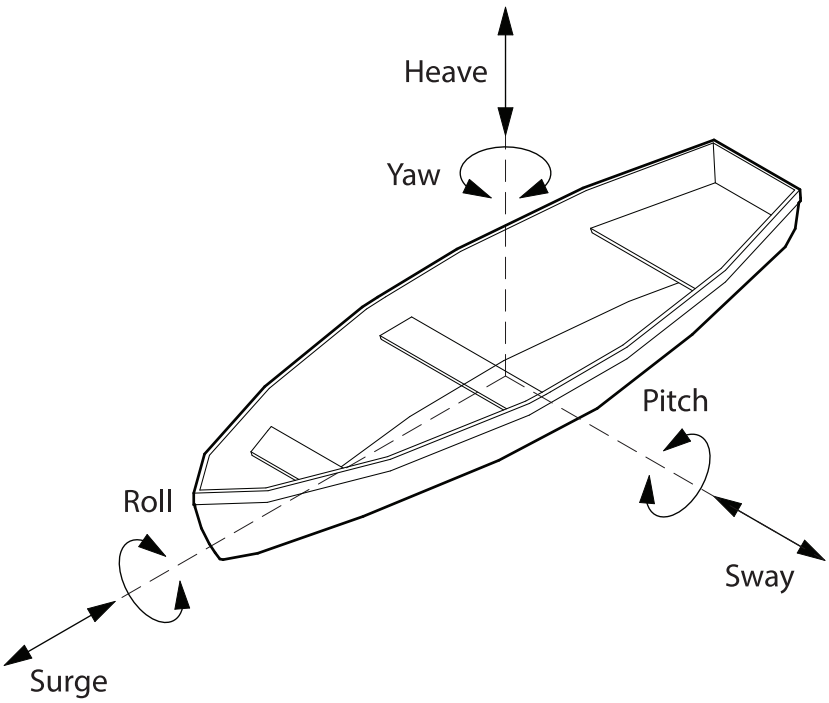


Figure 1.2: The six degrees of motion of a sonar system.

of drift, although the key point is the combined effect of the six types of motion on slant-range deviations. Corruption in reconstructed images due to undesired motion can be addressed to some extent by performing motion compensation, which requires navigation estimates of the actual path travelled.

A SAS system is typically equipped with an inertial navigation system (consisting of accelerometers and gyroscopes augmented with magnetometers, and Doppler velocity logs, etc.) used to estimate the position and orientation of the sonar. The sonar path can thus be taken into account when reconstructing the SAS image, reducing artefacts. Alternatively, the echo data itself can be used to estimate and correct the sonar motion using micronavigation and autofocus algorithms [Callow 2003]. It is also practical to use a combination of the data from both hardware-based estimates and data-driven estimates for further improvement.

1.9 Change detection and image registration

In repeat-pass sonar, the same scene is imaged two or more times, usually with the sonar path kept as similar as possible in both runs. A significant proportion of ongoing SAS research is focused towards repeat-pass applications such as change detection, where the purpose is to identify changes in the same scene between runs. Automatic change detection is of particular interest for mine hunting [Fandos 2012][Leier 2014], as well as other applications such as hydrography, ocean science, oil and gas exploration, and surveying of shipwrecks and underwater pipes [Griffiths et al. 1997][Dillon 2013].

It is an ongoing challenge to acquire high resolution SAS images with minimal artefacts due to the issues associated with navigation, and so the problem of detecting differences between these imperfect images is a more problematic and involved process. The first step of change detection is to *register* the repeat-pass images. Change detection is reliant on the images being aligned to a high accuracy, such that the advancement of change detection for practical applications is primarily contingent on improvements to image formation and image registration techniques. This thesis is dedicated to the topic of image registration using feature algorithms from the field of computer vision; the application of these techniques to sonar is a relatively recent endeavour with the supposed merit of faster estimation, though many aspects of the feasibility of feature-based registration are yet to be clarified.

1.10 Thesis outline and assumed knowledge

The main chapters that present novel contributions and findings are Chapters 5–7, for which the relevant background knowledge is established in Chapters 2–4.

Chapter 2 describes the phenomenon of speckle noise that affects sonar images, including models of speckle, measures of speckle (speckle contrast, coherence, and spatial coherence), the statistics of coherence, and sources of decorrelation.

Chapter 3 provides an introduction to local image features and the feature matching pipeline commonly used to perform registration in computer vision applications. This covers several concepts including feature detection, feature description, feature matching, the SIFT (Scale Invariant Feature Transform) and SURF (Speeded Up Robust Features) feature detectors/descriptors, robust estimation using Random Sample Consensus (RANSAC), and performance metrics.

Chapter 4 gives an overview of the problem of sonar image registration, including area-based methods and feature-based methods. A summary of related topics and a brief literature review of existing work is also given.

Chapter 5 presents a proof-of-concept feature-based pipeline for image registration via track registration, demonstrating the results for a repeat-pass simulated SAS scene.

Chapter 6 considers the effects of sinc-interpolated sub-pixel shifts and oversampling factor on the feature matching performance of SIFT based on randomly generated bland speckle images, highlighting the implications of these results.

Chapter 7 examines the relationship between feature matching performance and the coherence between an image pair using randomly generated correlated image pairs. A statistical model is proposed for the number of feature matches at a given coherence, and it is shown to accurately describe the performance of SIFT and SURF for both bland scenes and artifical ripple scenes.

Chapter 8 summarises the findings from this work and provides recommendations for future work.

The reader is presumed to have some level of familiarity with statistics and signal and image processing concepts. A prior background in sonar imaging and the use of features in computer vision is advantageous for contextual understanding but is not strictly necessary.

1.11 Publications

Material from Chapter 5 is based on work featured in the following publications:

> Wang, V. and Hayes, M. P. (2014). Image registration of simulated synthetic aperture sonar images using SIFT. In *Image and Vision Computing New Zealand*, pages 31–36. ACM.
>
> Wang, V. and Hayes, M. P. (2017b). Synthetic aperture sonar track registration using SIFT image correspondences. *IEEE Journal of Oceanic Engineering*, 42(4):901–913.

The material in Chapter 6 was presented in:

> Wang, V. and Hayes, M. (2017a). SIFT localisation accuracy on interpolated speckle images. In *Image and Vision Computing New Zealand*. IEEE.

Chapter 7 is based on the work from:

> Wang, V. and Hayes, M. P. (2016b). Analysis of feature matching performance on correlated speckle image pairs. In *OCEANS*. MTS/IEEE.
>
> Wang, V. and Hayes, M. (2016a). Modelling of feature matching performance on correlated speckle images. In *Image and Vision Computing New Zealand*. IEEE.

Chapter 2

Speckle and decorrelation

Coherent imaging systems such as sonar, radar, laser, and ultrasound exhibit speckle noise [Dainty 1984], which is an aspect-dependent random-like deterministic interference backscattering pattern resulting from the coherent summation of echoes from multiple independent scatterers in the scene [Goodman 1976]. ([Hunter 2006] provides a primer on the physics of acoustic wave propagation and acoustic scattering models.) Speckle noise has a granular appearance due to its multiplicative nature, which corrupts image quality by reducing contrast resolution, and poses a significant obstacle for reliable image interpretation by both humans and computers. Although speckle patterns have a random appearance, speckle is deterministic such that the same imaging system operating from the exact same location or path will observe the same speckle pattern. When an unchanged scene is imaged with slightly different positioning, the resulting speckle pattern is similar but differs according to the difference in positioning [Li and Goldstein 1990]. When the scene changes, the observed speckle pattern also changes.

The properties of speckle are relevant for several reasons. Images can be improved for interpretation purposes using despeckling filters [Lee et al. 1994] designed to reduce speckle while retaining details not caused by speckle [Lopes et al. 1993][Yu and Acton 2002]. Parts of a scene can be analysed according to the expected distribution of speckle statistics for different textures and surfaces. Multiple images of the same scene with similar speckle patterns can be combined to form a more accurate image where the variance of the speckle noise is reduced, as in multi-look processing [Huang and van Genderen 1997]. Although speckle noise is generally considered as a problem to overcome or minimise, the information from a speckle pattern can be useful in some applications. For example, a dynamic speckle pattern can be used to measure the temporal activity of an illuminated material, such as to monitor the drying of paint using laser speckle [Faccia et al. 2009].

In repeat-pass sonar, the same scene is imaged two or more times, usually with the sonar path kept as similar as possible across runs. The main repeat-pass sonar application is change detection, where the purpose is to identify changes in the scene

between two runs. The similarity between repeat-pass images is dependent on a number of factors, of which variations in the observed speckle patterns are one source of measured differences. The canonical measure of similarity between SAS images is *coherence*, which estimates the *correlation* (also a statistical measure) between the images. *Decorrelation* refers to decreased coherence (or correlation) between two sonar runs (or a pair of transducers) due to one or more of the categorical sources of change.

This chapter provides an overview of the background theory of speckle and decorrelation for a SAS system. Section 2.1 describes common statistical models of speckle noise. Section 2.2 establishes definitions of correlation and coherence in the context of random processes and explains their significance in signal processing. Section 2.3 describes how measures of speckle such as speckle contrast can be used to distinguish regions of a speckle image according to content. Section 2.4 briefly describes multilook processing, which can be used to reduce speckle noise. Section 2.5 introduces the statistics of coherence, providing context for the estimation of coherence. Section 2.6 presents the model of coherence factors for a SAS system, which includes five sources of decorrelation: decorrelation due to acoustic noise, a loss of coherence due to a footprint shift, baseline decorrelation, decorrelation due to processing noise, and temporal decorrelation.

2.1 Models of speckle

Although speckle is deterministic, it is implausible to model or predict exactly due to incomplete information about the scene elements and the reflective properties of the imaged surfaces at a sub-wavelength scale [Kuttikkad and Chellappa 2000]. Thus, it is useful to use statistical descriptions of speckle. A common model for speckle is known as *fully developed speckle*; it is common due to its simplicity and is accurate under a set of ideal conditions [Goodman 1975]:

- The resolution size of the sonar image is large with respect to the wavelength of the system such that there are a large number of scatterers within each resolution cell that contribute to the measured signal.
- The scatterers within a resolution cell are independent.
- No scatterer within the resolution cell is so reflective that it dominates the overall echo response.
- The phase of each scatterer is random and uniformly distributed over $[0, 2\pi)$.

When these conditions are met, speckle can be modelled by a random walk in the complex plane where each step is the echo response from a single scatterer within the given resolution cell, the resulting variable being the sum over all the individual

scatterers. The speckle value can be represented as the sum of a real component and an imaginary component using two real random variables X and Y:

$$Z = X + \mathrm{j}Y. \tag{2.1}$$

The assumptions for fully developed speckle invoke the central limit theorem, so that X and Y are independent zero-mean Gaussian-distributed variables with the same variance σ_X; equivalently, Z has a circular symmetric complex Gaussian distribution. The joint probability density function (PDF) of X and Y is the uncorrelated zero-mean bivariate Gaussian distribution:

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma_X^2} \exp\left(-\frac{x^2 + y^2}{2\sigma_X^2}\right). \tag{2.2}$$

The speckle magnitude $M = |Z| = \sqrt{X^2 + Y^2}$ is a Rayleigh-distributed random variable $M \sim \text{Rayleigh}(\sigma_X)$ with PDF

$$f_{|Z|}(z) = \frac{z}{\sigma_X^2} \exp\left(-\frac{z^2}{2\sigma_X^2}\right), \quad x \geq 0, \tag{2.3}$$

and the speckle intensity $I = |Z|^2 = X^2 + Y^2$ follows a negative exponential distribution $I \sim \text{Exp}(\frac{1}{\sigma_I})$ with

$$f_I(i) = \frac{1}{\sigma_I} \exp\left(-\frac{i}{\sigma_I}\right), \quad i \geq 0, \tag{2.4}$$

where $\sigma_I = 2\sigma_X^2$. The mean and variance of the speckle intensity are:

$$\mathbb{E}[I] = \sigma_I, \tag{2.5}$$

$$\text{Var}[I] = \sigma_I^2. \tag{2.6}$$

With SAS imaging, the mean intensity σ_I represents the reflectivity of the seafloor. It is equivalent to model speckle intensity as a normalised random variable that is multiplied by the scene reflectivity to give the observed sonar return, an approach taken in Section 7.1.

2.1.1 Other models

The Rayleigh speckle magnitude model is consistent with real measurements over homogeneous regions of a scene, especially when the spatial resolution is coarse. However, the model is less consistent at finer resolutions and over heterogeneous regions, where the assumptions of fully developed speckle are not met [Fortune 2005]. In practice, speckle distributions are often heavy tailed. Non-Rayleigh speckle can arise in uniform

regions due to mechanisms such as non-uniform fractal surface roughness of the seafloor and further scattering due to water-column turbulence [Lyons et al. 2010].

The K-distribution is a popular alternative to the Rayleigh distribution and arises from the assumption that the mean of the intensity of a resolution cell is gamma distributed. For normalised data with a mean power of one, the K-distribution model can be formulated with a single parameter, a shape parameter. The K-distribution is derived from a physical scattering process and reduces to the Rayleigh distribution as the shape parameter becomes infinite for homogeneous media. The K-distribution has been shown to be a suitable model for sonar returns and sonar speckle statistics [Jakeman and Pusey 1976][Dunlop 1997] and can be used to model the statistics of ripple scenes [Lyons et al. 2010].

Whereas the Rayleigh model assumes uniform reflectivity, a region consisting of the same content (such as a bland scene) may have non-uniform reflectivity due to the scene texture, which may account for variations in the reflectivity within each pixel. The log-normal distribution and the Weibull distribution are two popular models for fitting to real data of rough surfaces, although they are not based on the theoretical physics of scattering. Another empirical distribution is the Fisher distribution [Tison et al. 2003]. A survey of statistical models is given in [Gao 2010].

Second-order statistics of speckle in the context of SAS imaging such as the speckle spectrum, speckle correlation in the along-track and across-track directions, *speckle size* [Dainty 1984], and the effect of sampling are detailed in [Fortune et al. 2004].

2.2 Correlation and coherence

Correlation and coherence are two terms that are used in many different fields, sometimes to mean similar/related ideas and sometimes not. Therefore, it is appropriate to clarify their mathematical meanings as mathematical functions or measures. This section establishes correlation for random processes, correlation for signals, spatial correlation, coherence, and the common ideas behind these definitions. A brief background on their mathematical contexts is also given. Some common but unnecessary definitions such as correlation for random variables (the Pearson correlation coefficient) and phase coherence are omitted. The term correlation often refers to the cross-correlation and has the same connotation in this thesis.

There are multiple definitions of correlation in use in signal processing literature. For example, across the different contexts (random variables, random processes, signals, images, etc.), the correlation function may be defined as centralised or uncentralised, normalised or unnormalised. (The uncentralised version of correlation gives equivalent results when the input signals are zero-mean, as is often assumed to be the case.) In addition, the factor that conjugation is performed on in the complex generalisation of correlation and covariance can be arbitrarily chosen as the first factor or the second.

(Conjugation on the first factor is common for signal processing, as opposed to the second factor in the field of statistics.) Lastly, a positive time delay τ can be arbitrarily defined as indicating either a leading or lagging relationship between two signals. The differences between these choices are mostly a matter of convention and do not impair the general ideas of interpretation.

Introductory texts generally cover a limited subset of the concept of correlation while adopting distinct definitions, notations, and conventions, such that any collation of definitions cannot be consistent with any single textbook. This thesis presents its own unique set of definitions that is both self-consistent and appropriate in the context of sonar image processing. Much of this content was adapted from and resembles the notation of [Gray and Davisson 2004]. These formal definitions seem necessary since most papers on sonar imaging discuss correlation without including or referencing any explicit definition. However, it is impractical to include all the possible variants of autocovariance and autocorrelation, and any alternative and implied definitions should be easily deducible. For example, the relationship between autocovariance and autocorrelation is roughly the same in all the various domain spaces. The relationships between cross-covariance and autocovariance, and between cross-correlation and autocorrelation are also consistent. Although many texts only define the autocovariance and autocorrelation, the cross-covariance and cross-correlation functions can be considered more general versions of these.

2.2.1 Random processes

Let X be a random variable defined on a probability space $(\Omega, \mathcal{F}, P)$ consisting of sample space Ω, event space $\mathcal{F}$, and probability measure P. The expected value of X is defined as the Lebesgue integral

$$\mathbb{E}[X] = \int_{\Omega} X(\omega)\,\mathrm{d}P(w). \tag{2.7}$$

Note that the integral (and thus the expectation) may not exist, in which case X is said to have infinite expectation.

If $F_X(x) = \mathrm{P}(X \leq x)$ is the cumulative distribution function of X, then

$$\mathbb{E}[X] = \int x\,\mathrm{d}F_X(x) = \begin{cases} \displaystyle\sum_x x\,p_X(x) & \text{if } X \text{ is discrete and } p_X(x) \text{ is the probability mass function of } X; \\ \displaystyle\int x\,f_X(x)\,\mathrm{d}x & \text{if } X \text{ is continuous and } f_X(x) \text{ is the probability density function of } X. \end{cases} \tag{2.8}$$

For two random variables X and Y, whether real-valued or complex-valued, their

covariance can be defined as

$$\operatorname{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])^*(Y - \mathbb{E}[Y])]. \tag{2.9}$$

Note that for complex variables, the conjugation applied to the first factor ensures that the imaginary components of the values also contribute to the similarity value in terms of the phase between the two factors [Therrien 1999]. Now let there be a random process (also known as a stochastic process), i.e., an indexed family of random variables $\{X_t;\ t \in \mathrm{T}\}$ defined on a common probability space $(\Omega, \mathcal{F}, P)$ and where T may be discrete (e.g., $\mathrm{T} = \mathbb{Z}_+$ or $\mathrm{T} = \mathbb{Z}$) or continuous (e.g., $\mathrm{T} = \mathbb{R}$). The autocovariance function is defined by

$$C_X(t, s) = \operatorname{cov}(X_t, X_s) = \mathbb{E}[(X_t - \mathbb{E}[X_t])^*(X_s - \mathbb{E}[X_s])] = \mathbb{E}[X_t{}^* X_s] - \mathbb{E}[X_t{}^*]\mathbb{E}[X_s]. \tag{2.10}$$

Thus, the autocovariance function is the covariance of all possible pairs of samples. More generally, the covariance function for two random processes $\{X_t\}$ and $\{Y_t\}$; $t \in \mathrm{T}$ is defined as

$$C_{XY}(t, s) = \operatorname{cov}(X_t, Y_s) = \mathbb{E}[(X_t - \mathbb{E}[X_t])^*(Y_s - \mathbb{E}[Y_s])] = \mathbb{E}[X_t{}^* Y_s] - \mathbb{E}[X_t{}^*]\mathbb{E}[Y_s]. \tag{2.11}$$

Autocovariance refers to the case where the samples come from a single process, whereas the term *cross-covariance* is used when the samples are taken from two processes. Thus, the notation $C_X(t, s)$ for autocovariance can be considered a shortened form of $C_{XX}(t, s)$. (However, in some texts the autocovariance function is simply referred to as the covariance function.)

2.2.2 Wide-sense stationary random processes

A random process $\{X_t\}$ is said to be wide-sense stationary (WSS) or weakly stationary if its mean and autocovariance are time-invariant. Formally, the mean function does not depend on time if

$$\mathbb{E}[X_t] = \mathbb{E}[X_{t+\tau}] = \mu_X \text{ for all } t, \tau \in \mathrm{T}. \tag{2.12}$$

The autocovariance does not depend on time if $C_X(t, s)$ depends on t and s only through the difference $s - t$, i.e.:

$$C_X(t, s) = C_X(t + \tau, s + \tau) \text{ for all } t, s, \tau \text{ where } s, s + \tau, t, t + \tau \in \mathrm{T}. \tag{2.13}$$

Thus, the autocovariance of a wide-sense stationary process can be simplified and written as $C_X(t, s) = C_X(0, s - t) = C_X(\tau)$, where $\tau = s - t$ is analogous to the time delay of a second signal relative to the first. Since the autocovariance of a WSS process

is time-invariant, the variance $\sigma_{X_t}^2 = C_X(t,t) = \sigma_X^2$ is also time-invariant.

Two random processes $\{X_t\}$ and $\{Y_t\}$ are jointly WSS if each is WSS and additionally their cross-covariance $C_{XY}(t,s)$ is time-invariant (i.e., it depends only on $\tau = t - s$). Thus, the cross-covariance function for two jointly WSS functions can be written as

$$C_{XY}(\tau) = C_{XY}(t,s) = \mathrm{cov}(X_t, Y_s) = \mathbb{E}[{X_t}^* Y_s] - {\mu_X}^* \mu_Y = \mathbb{E}[{X_{t-s}}^* Y_0] - {\mu_X}^* \mu_Y. \tag{2.14}$$

The (normalised) correlation function for two jointly WSS random processes is defined as

$$\begin{aligned} \rho_{XY}(\tau) &= \frac{1}{\sigma_X \sigma_Y} \mathbb{E}[(X_\tau - \mu_X)^* (Y_0 - \mu_Y)] \\ &= \frac{1}{\sigma_X \sigma_Y} (\mathbb{E}[{X_\tau}^* Y_0] - {\mu_X}^* \mu_Y) = \frac{C_{XY}(\tau)}{\sigma_X \sigma_Y}. \end{aligned} \tag{2.15}$$

2.2.3 Ergodicity

A stationary process has a constant mean, as in (2.12). This mean function measures the ensemble mean at a given point in time. Another concept of a mean is the *time average*, which relates to a notable statistical property called *ergodicity*.

Formally, an event F is said to be τ-*invariant* if $\{x_t;\ t \in \mathrm{T}\} \in F$ implies that also $\{x_{t+\tau};\ t \in \mathrm{T}\} \in F$, i.e., if a given sequence or waveform is in F, then the same sequence or waveform shifted by τ is also in F. A random process $\{X_t;\ t \in \mathrm{T}\}$ is *ergodic* if for any value of τ, all τ-*invariant* events F have probability zero or one [Gray and Davisson 2004]. In the case of a discrete time process, it is sufficient to consider only $\tau = 1$. The implications of ergodicity can be seen in the pointwise (or strong) ergodic theorem [Birkhoff 1931], which can be stated as follows.

Given a discrete time stationary random process $\{X_n;\ n \in \mathbb{Z}\}$ with finite expectation $E(X_n) = \mu_X$, there exists a variable $\hat{X}$ such that:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{n=0}^{\infty} X_n = \hat{X} \text{ with probability one;} \tag{2.16}$$

that is, the limit exists and the value of the limit is equal to $\hat{X}$ with probability one. If the process is also ergodic, then the random variable $\hat{X}$ is simply a constant, μ_X:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{n=0}^{\infty} X_n = \mu_X \text{ with probability one.} \tag{2.17}$$

A continuous time version of the theorem also exists in the form:

$$\lim_{T\to\infty} \frac{1}{T} \int_0^\infty X(t)\,\mathrm{d}t = \hat{X} \text{ with probability one,} \tag{2.18}$$

however, special conditions are required to formally ensure the existence of time-average integrals.

For a process that is both stationary and ergodic, time averages converge to the same value as the ensemble mean with probability one. This is useful because it implies that a time average of a signal can be used to estimate the ensemble mean of a stationary ergodic process. If a process is stationary but not ergodic, time averages still converge, but not necessarily to the ensemble mean. On the other hand, ergodicity implies stationarity.

2.2.4 Correlation of signals

In practice, a limited number of realisations (sometimes only one realisation) of a random process can be observed. Given two continuous-time signals $x(t)$ and $y(t)$ that are the realisations of two jointly WSS random processes $\{X_t\}$ and $\{Y_t\}$, the cross-correlation function can be defined as

$$\hat{\rho}_{xy}(\tau) = \frac{1}{\hat{\sigma}_x \hat{\sigma}_y} \int_{-\infty}^{\infty} (x(t) - \mu_x)^* (y(t) - \mu_y)\,\mathrm{d}t. \tag{2.19}$$

This function is normalised by the means (μ_x, μ_y) and estimated standard deviations ($\hat{\sigma}_x$, $\hat{\sigma}_y$) of $x(t)$ and $y(t)$ and is not the most common definition. The intuitive definition of the sample standard deviation of a continuous function is supposed. This formalisation resembles (2.15) and can be considered an estimate of the correlation $\rho_{XY}(\tau)$. Correlation of deterministic data can be referred to as *deterministic correlation*. However, the estimates of mean and standard deviation may not be accurate, since the signals are only a single realisation of the underlying joint random process. If the underlying jointly WSS processes are discrete, the corresponding definition of the correlation function for $x[t]$ and $y[t]$ is

$$\hat{\rho}_{xy}(\tau) = \frac{1}{\hat{\sigma}_x \hat{\sigma}_y} \sum_{n=-\infty}^{\infty} (x[n] - \mu_x)^* (y[n+\tau] - \mu_y). \tag{2.20}$$

Note that a realisation of a discrete random process is also referred to as a *time series*.

It is more convenient to consider that a measured signal of interest will consist of a finite number of contiguous samples, as well as containing a finite amount of "energy". Let the time series be defined as $\{x_0, x_1, \ldots, x_{N-1}\}$ and $\{y_0, y_1, \ldots, y_{N-1}\}$, each consisting of N samples. (If one sequence is shorter than the other, the shorter one can be padded with zeroes on the end to obtain the same length N.) The following

formulation of the correlation between two time series $x[t]$ and $y[t]$ for an integer delay, τ, is called the *sample cross-correlation* and is defined by:

$$\hat{\rho}_{xy}(\tau) = \begin{cases} \dfrac{1}{\hat{\sigma_x}\hat{\sigma_y}} \displaystyle\sum_{t=0}^{N-\tau-1} (x[t] - \mu_x)^*(y[t+\tau] - \mu_y) & \text{for } 0 \leq \tau \leq N-1; \\ \hat{\rho}_{yx}(-\tau) & \text{for } -(N-1) \leq \tau \leq -1. \end{cases} \tag{2.21}$$

The cross-correlation functions defined in (2.19)–(2.21) can be used to estimate the correlation between two (supposedly) jointly WSS random processes. Joint stationarity ensures that these estimates converge to a fixed value, however, joint ergodicity is required to ensure that the asymptotic estimate is actually equal to the true correlation. Joint ergodicity represents the case where time averages converge to the same value as the ensemble mean, and thus estimation can be performed from a single realisation of each random process. Estimation bias can be reduced by replacing the sample means (μ_x and μ_y) and sample standard deviations (σ_x and σ_y) by the true means (μ_X and μ_Y) and true standard deviations (σ_X and σ_Y) where available. In sonar and radar, the assumption of ergodicity corresponds to the idea that the spatial average of many scatterers within a resolution cell is equal to the ensemble average of a single scatterer [Birkhoff 1931].

2.2.5 Image correlation

A stochastic process can be generalised as a *random field*, where the underlying parameter can be a multidimensional vector representing a point in any space, often an n-dimensional Euclidean space. (A formal introduction to random fields can be found in [Abrahamsen 1997].) A 2D image, for example, can be treated as a realisation of a random field defined on a 2D space with discrete coordinates. The continuous space can also be treated as discrete, especially in the case of interpolating spatially correlated values. Fully developed speckle can be modelled as a Gaussian random field [Dainty 1980].

Discrete 2D sample cross-correlation (often referred to as image correlation) between two realisations $x[m,n]$ and $y[m,n]$ of the same size $M \times N$ can be defined as:

$$\hat{\rho}_{xy}(\nu,\tau) = \frac{\displaystyle\sum_{m=0}^{M-\nu-1} \sum_{n=0}^{N-\tau-1} (x[m,n] - \bar{x})^* \, (y[m+\nu, n+\tau] - \bar{y})}{\sqrt{\left(\displaystyle\sum_{m=0}^{M-1} \sum_{n=0}^{N-1} |x[m,n] - \bar{x}|^2 \right) \left(\displaystyle\sum_{m=0}^{M-1} \sum_{n=0}^{N-1} |y[m,n] - \bar{y}|)^2 \right)}}, \tag{2.22}$$

for $0 \leq \nu \leq M-1$ and $0 \leq \tau \leq N-1$, where ν and τ are integers, and $\bar{x}$ and $\bar{y}$ are the sample mean values of the images. If the true mean values of the underlying random

fields are known, these can be used instead of the sample means.

In case of negative offsets or delays, the full definition is:

$$\hat{\rho}_{xy}(\nu,\tau) = \begin{cases} (2.22) & \text{for } 0 \;\leq\; \nu \;\leq M-1 \text{ and } 0 \leq \tau \leq N-1; \\ \hat{\rho}_{y'x'}(\nu,-\tau) & \text{for } 0 \;\leq\; \nu \;\leq M-1 \text{ and } -(N-1) \leq \tau \leq -1; \\ \hat{\rho}_{x'y'}(-\nu,\tau) & \text{for } -(M-1) \leq \nu \leq -1 \text{ and } 0 \leq \tau \leq N-1; \\ \hat{\rho}_{yx}(-\nu,-\tau) & \text{for } -(M-1) \leq \nu \leq -1 \text{ and } -(N-1) \leq \tau \leq -1, \end{cases} \tag{2.23}$$

where $'$ denotes vertical flipping of an image such that $x'[m,n] = x[M-1-m,n]$ for $0 \leq m \leq M-1$.

With data that can be modelled by zero-mean random fields (which includes raw sonar data), the cross-correlation estimate can be improved and simplified to:

$$\hat{\rho}_{xy}(\nu,\tau) = \frac{\sum_{m=0}^{M-\nu-1} \sum_{n=0}^{N-\tau-1} x^*[m,n]\, y[m+\nu,n+\tau]}{\sqrt{\left(\sum_{m=0}^{M-1} \sum_{n=0}^{N-1} |x[m,n]|^2\right)\left(\sum_{m=0}^{M-1} \sum_{n=0}^{N-1} |y[m,n]|^2\right)}}. \tag{2.24}$$

Image correlation is a spatial correlation, so it only estimates cross-correlation in the ensemble sense when the underlying random fields are ergodic.

With template matching, the goal is to detect any instances of the template image appearing in a larger output image [Lewis 1995]. Likely instances are indicated by local maxima close to a value of one in the output matrix of image correlation. Although the definition (2.22) is specified in terms of images of the same size, it is trivial to implement this function without any padding.

2.2.6 Coherence

The complex coherence of two jointly zero-mean WSS random processes $\{X_t\}$ and $\{Y_t\}$ is defined as the correlation at zero delay [Touzi et al. 1999] [Born and Wolf 1999]:

$$\rho = \rho_{XY}(0) = \frac{\mathbb{E}[X^*Y]}{\sqrt{\mathbb{E}\left[|X|^2\right]}\sqrt{\mathbb{E}\left[|Y|^2\right]}}. \tag{2.25}$$

Due to stationarity, these expectation values can be calculated at any time offset. If the processes are not jointly ergodic, then these ensemble expectations must be calculated from all the realisations at a given time offset (or in practice, estimated from multiple realisations). If the processes are jointly ergodic, then the coherence can be estimated using temporal/spatial correlation. The *sample coherence* of complex

signals x and y is defined as:

$$\hat{\rho} = \hat{\rho}_{xy}(0) = \frac{\sum_{i=1}^{N} x_i^* y_i}{\sqrt{\sum_{i=1}^{N} |x_i|^2} \sqrt{\sum_{i=1}^{N} |y_i|^2}}. \tag{2.26}$$

Here, the expectations are estimated from N measurements that can be drawn from any joint realisation and time offset. This formula also applies to other geometries including the case of 2D image correlation of realisations of random fields, where the sample coherence $\hat{\rho}$ can be calculated as $\hat{\rho}_{xy}(0,0)$ from (2.24). However, the assumption of ergodicity is often unrealistic or unnecessary for a whole image but can instead be applied to a local image patch. This idea is the basis of *windowed* correlation [Boker et al. 2002], where the coherence of supposedly matching pixels across two images is estimated as the zero-delay image cross-correlation over two rectangular subimages (of a designated window size) encompassing the matching pixels centrally.

The coherence estimator is asymptotically unbiased (due to ergodicity) [Touzi et al. 1999] but generally tends to overestimate the true coherence. The statistics of coherence estimation are covered in Section 2.5.

2.2.7 Interpretation of correlation and covariance

The covariance between two random variables is a measure of their joint variability. If greater values of one variable tend to be observed in tandem with greater values of the other and a similar trend is also present for the lower values, then the covariance is positive. If the variables tend to correspond in an opposite manner, such that greater values of one variable tend to be observed with lesser values of the other, then the covariance is negative. If the covariance is zero, the variables are said to be uncorrelated.

Covariance can be interpreted as a measure of the linear dependence between two variables. The better the linear relationship, the higher the magnitude of the covariance. A large positive value indicates a positive linear relationship, whereas a large negative value indicates a negative linear relationship. However, the interpretation of a covariance value in terms of linear dependence is not always clear, since scaling of one variable by a constant factor results in scaling of the covariance by the same factor. The Pearson correlation coefficient [Benesty et al. 2009] (often referred to as just the correlation coefficient) is defined as a normalised version of covariance in order to address this problem. A correlation coefficient (and also covariance) of zero implies no linear relationship (but not necessarily independence), whereas a correlation of ± 1 indicates that the two variables can be perfectly described by a linear relationship. Independent variables always have a covariance of zero.

The interpretations of correlation and covariance, whether defined for random processes, signals, or images, are similar to the given interpretation for random variables. For example, in each version of cross-correlation, the value of the function for a given set of arguments is always (in an informal sense) a cross-correlation coefficient of two random variables. For random processes, the correlation and covariance functions yield the correlation coefficient and covariance values (respectively) for the random variables at the two specified time offsets. When the two random processes are jointly WSS, the interpretation does not change, but the functions reduce to being dependent only on the delay between the two times. When two discrete random processes are jointly ergodic (and therefore also jointly stationary), one set of time series realisations from the two processes can be used to estimate the true correlation and covariance without additional knowledge of the two underlying processes. In the context of signal processing, if for two signals the correlation $\rho_{xy}(\tau)$ is one or close to one at $\tau = s$, then under specific circumstances it may be likely that the underlying processes are equivalent, subject to a time delay s and some scale factor. A common scenario is of one signal being a detected echo of the other but with random noise. Correlation can be used to detect the presence of a signal [Poor 1994][Kassam and Thomas 1988] or estimate its time delay [Azaria and Hertz 1984][Tugnait 1993][Jacovitti and Scarano 1993].

2.2.8 Relation to matched filtering and convolution

Using time-domain correlation to detect a known waveform is equivalent to using the *matched filter* for that given signal. The matched filter is the only optimal linear filter for detecting a signal in the presence of (additive) random white noise, such as Gaussian noise [Turin 1960]. Specifically, the matched filter is optimal in the sense of producing a detection peak higher than the residual noise level (in terms of SNR) compared to any other linear filter [Vijaya Kumar and Hassebrook 1990][Spencer 2010]. It also provides the best point estimates for localisation, as finding the peak of a match filtered output corresponds to maximum likelihood estimation of delay under additive Gaussian noise [Röver 2011]. For template matching, the unnormalised correlation is motivated by the sum of squared differences or squared Euclidean distance [Lewis 1995], with the normalised cross-correlation being more robust to the template as well as changes in intensity [Trucco and Verri 1998].

Maximum likelihood estimation of delay can be adapted to non-white uncorrelated noise using the generalised correlation method [Knapp and Carter 1976][Fortune et al. 2004]. Matched filtering can be performed by correlating the output signal with the template signal, or alternatively, using convolution.

Convolution is similar to correlation and is defined for continuous functions as:

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau) g(t - \tau) \, \mathrm{d}\tau, \tag{2.27}$$

and for discrete functions as:

$$(f * g)[n] = \sum_{m=-\infty}^{\infty} f[m]g[n-m]. \tag{2.28}$$

In this context, the (unnormalised) discrete correlation is defined as

$$(f \star g)[n] = \sum_{m=-\infty}^{\infty} f^*[m]g[m+n], \tag{2.29}$$

where convolution and correlation are related by:

$$f(t) \star g(t) = f^*(-t) * g(t). \tag{2.30}$$

Computing the convolution in the Fourier domain using the FFT algorithm is faster than the naïve time-domain computation except for small input sizes. Correlation (and matched filtering) can be computed using the above relation by conjugating the first signal, as the Fourier transform is conjugate symmetric.

2.2.9 Summary

Correlation is a statistical measure of similarity with many closely related definitions under different contexts. The unnormalised definitions are appropriate for peak detection and delay estimation due to its connection with matched filtering and its optimality with signals corrupted by additive Gaussian noise, whereas the normalised versions are suitable for template matching, spatial coherence estimation, and measuring degree of similarity. Whereas correlation measures similarity with respect to the offset between two images, coherence measures the point-wise similarity across two zero-lag images. Thus, although coherence is equivalent to zero-lag correlation, calculating a correlation output matrix (used to determine delay or offset) is distinct from computing a coherence image (used to gauge localised changes or differences in the image).

2.3 Measuring speckle

For the fully developed speckle model, the speckle distributions apply to the ensemble magnitude or intensity values of a single point in the scene over all possible realisations. The distributions do not describe the distribution of speckle values over multiple points in the scene, unless the underlying random field representing the scene is ergodic. (In this case, ergodicity implies that any resolution cell within the scene or part of the scene satisfies the assumptions of fully developed speckle, which is equivalent to each pixel being independently and identically distributed.) Although scenes are not truly ergodic in practice, fully developed speckle can be used to approximate the distribution

of a small number of pixel values from a single speckle realisation if the expected value of each pixel (i.e., the reflectivity) is roughly uniform over the area [Fortune 2005] and the image consists of independent speckle values (i.e., the image is not oversampled in terms of the relationship between a pixel and the resolution size).

Figure 2.1 shows the distribution of sample pixel values from a SAS image of a patch of bland seafloor, plotted alongside the expected Gaussian distribution (for the magnitudes of the separate real and imaginary components of a speckle value), Rayleigh distribution (for speckle magnitude), negative exponential distribution (for speckle intensity), and uniform distribution (for phase) of fully developed speckle. The data somewhat conforms to the model; for limited sample sizes, the fully developed speckle model may be a good approximation even for scenes that are non-ergodic.

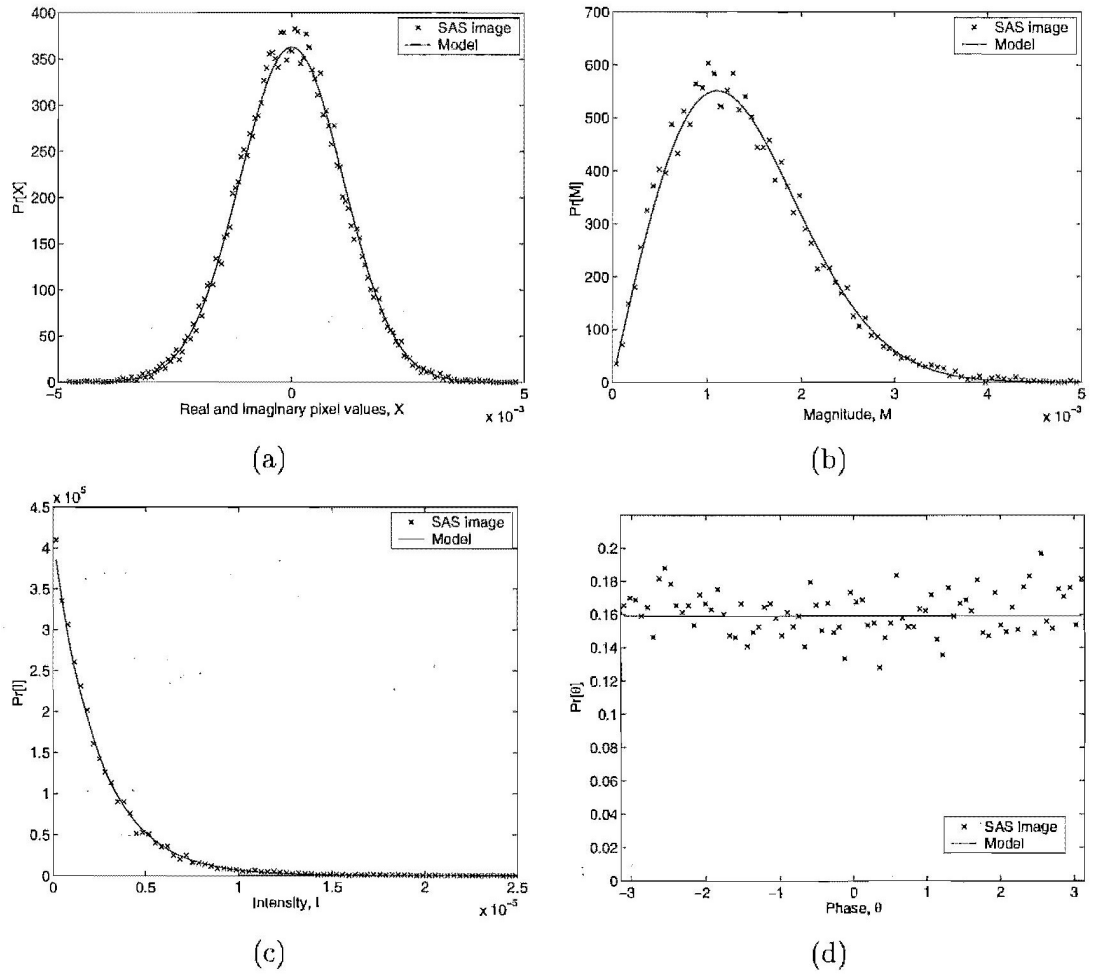


Figure 2.1: Models of the expected distributions for fully developed speckle compared to sample values from a bland seafloor SAS image. (a) Real and imaginary component values of speckle. (b) Pixel magnitudes. (c) Pixel intensities. (d) Pixel phase. Reprinted, with permission, from [Fortune 2005].

2.3.1 Speckle contrast

Speckle contrast is a common measure of the level of speckle present within an image. It is defined as the ratio of the standard deviation to the mean of the intensity of an area in the image [Goodman 1976]:

$$C = \frac{\mathrm{Std}[I]}{\mathbb{E}[I]}. \tag{2.31}$$

From (2.5) and (2.6), $C = 1$ holds for an exponentially distributed random variable for speckle intensity, I. Therefore, the speckle contrast is expected to be unity for regions that are consistent with the assumptions for fully developed speckle, whereas the contrast is expected to differ in regions where the assumptions are not met. A low contrast indicates relatively constant values over a region, whereas a contrast greater than one indicates a higher degree of variation than can be explained by fully developed speckle.

2.3.2 Other measures and applications

Scintillation and lacunarity are two measures of variation that are equivalent to each other and appear in sonar, radar, astronomy, and other contexts. As described in [Bonnett 2017], speckle contrast is equivalent to the square root of the scintillation index and lacunarity, and thus they encapsulate the same concept. Speckle contrast, scintillation, and lacunarity can be computed using spatial windows of different sizes. These contrast measures can be used for classification of textures. With SAS imagery, contrast can be used to detect objects within sand ripples and categorise the nature of seafloor patches [Nelson and Kingsbury 2012][Nelson and Krylov 2014][Williams 2015]. Classification of land use has also been performed in SAR [Dekker 2003]. Speckle contrast is an ensemble statistic, and can be estimated using spatial windows, temporal variations, or spatiotemporal methods [Boas and Dunn 2010]. Figure 2.2 shows a laser speckle image of a small area of a rat brain, where the speckle contrast is estimated using a spatial window to reveal blood vessels in the image. Regions with blood flow yield a lower contrast because of blurring of the speckle pattern, which occurs due to the integration time or exposure time of the camera [Boas and Dunn 2010].

2.4 Multi-look processing

Speckle noise can be reduced using *multi-look* techniques. In one form of multi-look processing the Doppler spectrum of the return spectrum is divided into N parts, effectively dividing the synthetic aperture, with each subset forming a separate image [Lu and Dzurisin 2014][Myers et al. 2017]. The multi-look image is formed from the incoherent average of these N images, where both the speckle and spatial resolution are

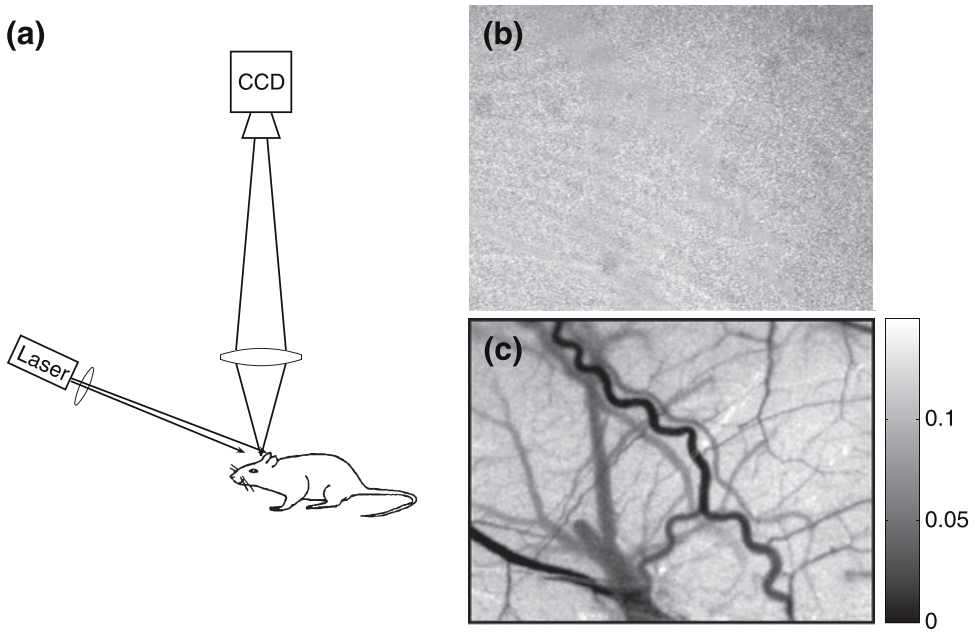


Figure 2.2: (a) Laser speckle contrast imaging (LSCI) setup with a laser diode and CCD (charge-coupled device) camera. (b) Raw speckle image of the rat cerebral cortex, taken through a thinned skull. (c) Regions of blood flow can be distinguished in the speckle contrast image, estimated using a sliding window.
From [Dunn 2012]. Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature. *Annals of Biomedical Engineering.* Laser Speckle Contrast Imaging of Cerebral Blood Flow. Andrew K. Dunn. © 2011.

reduced by a factor of N [Pohl and van Genderen 2016]. Another multi-look method refers to the spatial averaging of adjacent pixels of the single-look interferogram [Huang and van Genderen 1997][Lu and Dzurisin 2014]. Multi-look speckle values can be modelled by a gamma distribution based on the effective number of looks [Henderson and Lewis 1998]. Multi-look techniques are common in SAR but not SAS due to the use of higher frequencies and bandwidth and the relative accuracy/stability of repeat-pass navigation with aerial and satellite radar. Multi-look for SAS speckle reduction is considered in [Fortune et al. 2003] and is important for interferometric synthetic aperture sonar (InSAS) to reduce phase variance.

2.5 Statistics of coherence

Suppose there is a coherence, ρ, between two speckle values represented by random variables X and Y. (Note that the formula for the coherence between variables is virtually the same as in (2.25). ρ is known as the *complex degree of coherence* and the coherence magnitude $D = |\rho|$ is the *degree of coherence*, where $\hat{D} = |\hat{\rho}|$ is the maximum likelihood estimator of the degree of coherence. With SAS imaging, it is generally impractical to obtain enough realisations or *looks* to form a good estimate

of the coherence due to the reduction in resolution when using multi-look processing. Furthermore, with the deterministic behaviour of speckle, multi-pass imagery will not provide random (independent) realisations. Therefore, it is necessary to include the assumption that any corresponding pair of pixel values within a larger image subregion is characterised by the same coherence ρ. This allows sample coherence to be estimated over a window of pixels rather than over multiple (unattainable) realisations. For the case of two images governed by circular (complex) Gaussian statistics, the PDF of $\hat{D}$ is [Touzi and Lopes 1996]

$$f_{\hat{D}}(d \mid D) = 2(N-1)(1-D^2)^N d(1-d^2)^{N-2}\,{}_2F_1(N,N;1;D^2d^2), \quad 0 \le d \le 1, \tag{2.32}$$

where $N > 2$ is the number of independent samples or effective number of looks, $D \neq 1$, and ${}_pF_q$ is the generalised hypergeometric function. Note that the PDF is independent of the reflectivities, i.e., the expected intensities of X and Y. The kth moment of $\hat{D}$ is given by [Touzi et al. 1999]

$$\mathbb{E}[\hat{D}^k] = (1-D^2)^N \frac{\Gamma(N)\Gamma(1+k/2)}{\Gamma(N+k/2)}\,{}_3F_2(1+k/2,N,N;N+k/2,1;D^2), \quad D \neq 1, \tag{2.33}$$

and from this the first moment or mean of $\hat{D}$ can be calculated as

$$\mathbb{E}[\hat{D}] = (1-D^2)^N \frac{\Gamma(N)\Gamma(1.5)}{\Gamma(N+0.5)}\,{}_3F_2(1.5,N,N;N+0.5,1;D^2), \quad D \neq 1. \tag{2.34}$$

The expected value $\mathbb{E}[\hat{D}]$ converges to D as N approaches infinity but otherwise exceeds D. $\hat{D}$ is thus a biased (over)estimator of D, with the bias decreasing with more independent samples. Figure 2.3 shows the distribution of the estimated degree of coherence for different values of D with a fixed number of independent scatters, $N = 16$. Figure 2.4a shows the relationship between D and the expected value of $\hat{D}$ for different values of N, highlighting the larger bias for smaller values of coherence. The variance of the estimated degree of coherence is given by $\mathrm{Var}[\hat{D}] = \mathbb{E}[\hat{D}^2] - \mathbb{E}[\hat{D}]^2$; this variance with respect to D is plotted for different values of N in Figure 2.4b. The statistics of the phase (or argument) of the estimated complex degree of coherence is relevant for InSAS and is considered in [Touzi and Lopes 1996].

The above statistics are based on the assumption of a Gaussian scene. For a K-distributed scene, the number of scatterers per resolution cell fluctuates but is characterised by an overall average, with the overall statistics of coherence being identical to the Gaussian case [Yueh et al. 1989][Joughin et al. 1994].

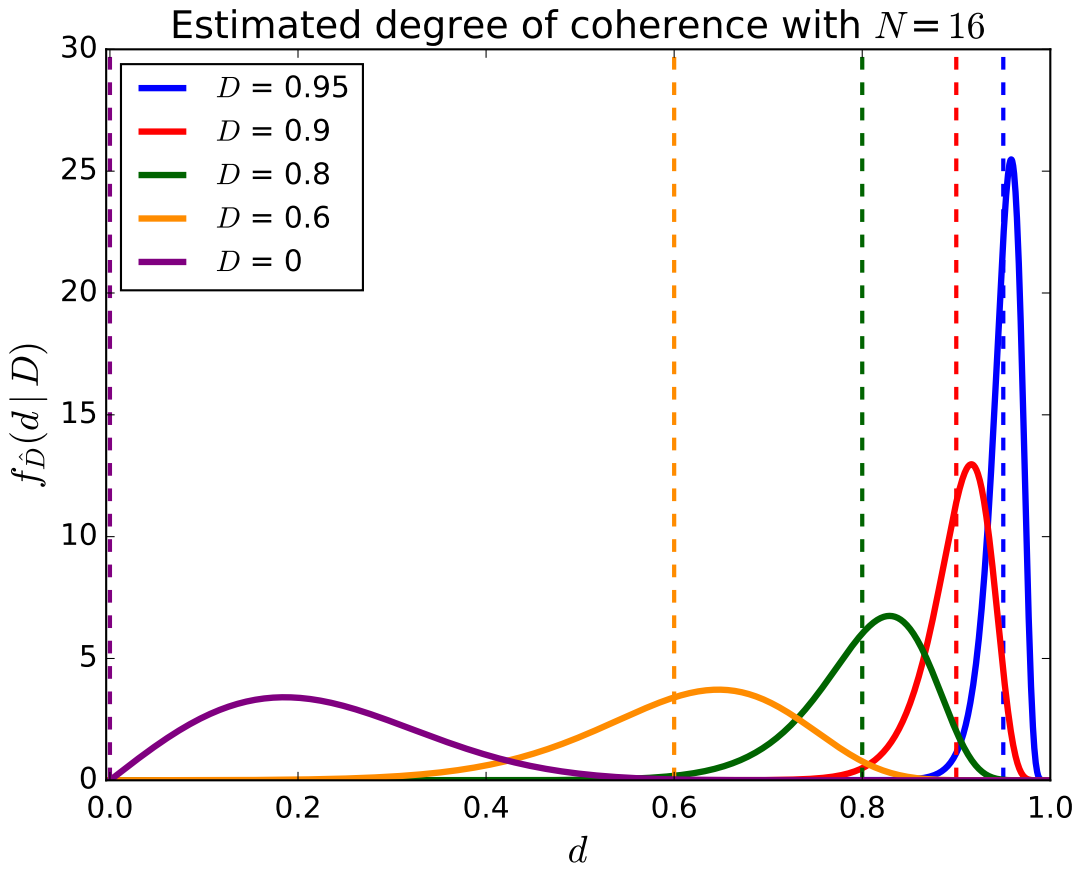


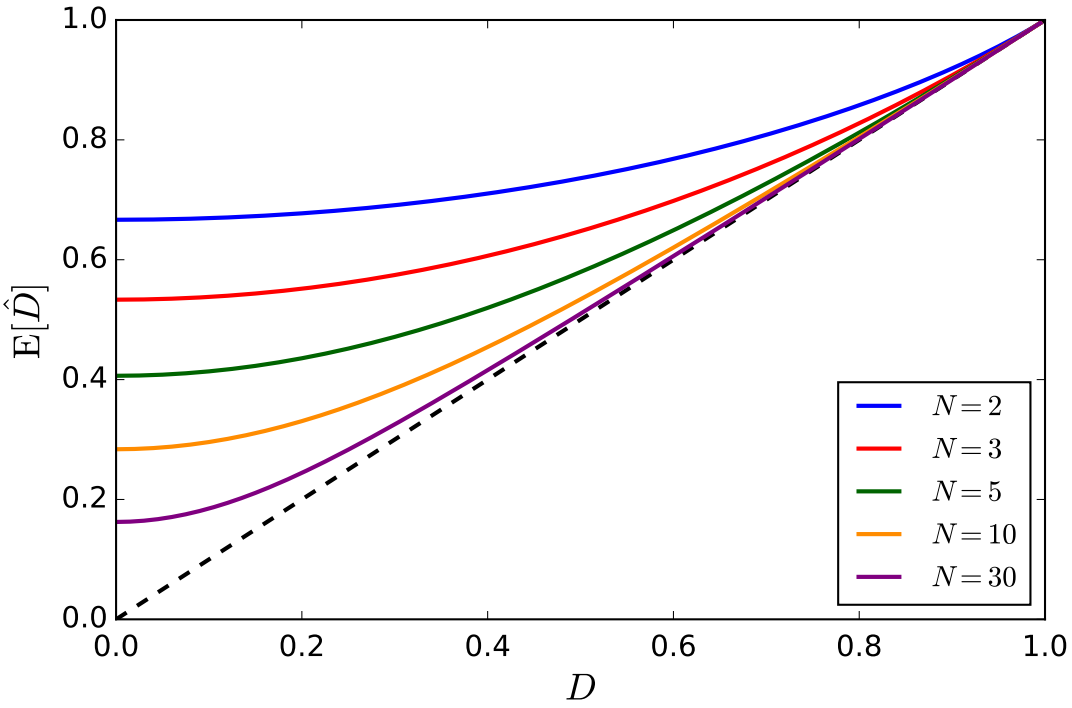
Figure 2.3: PDFs of the estimated degree of coherence with $N = 16$ looks for different values of the true coherence, D. The dashed lines indicate the true coherence values to highlight bias and spread.

2.6 Coherence factors in a SAS system

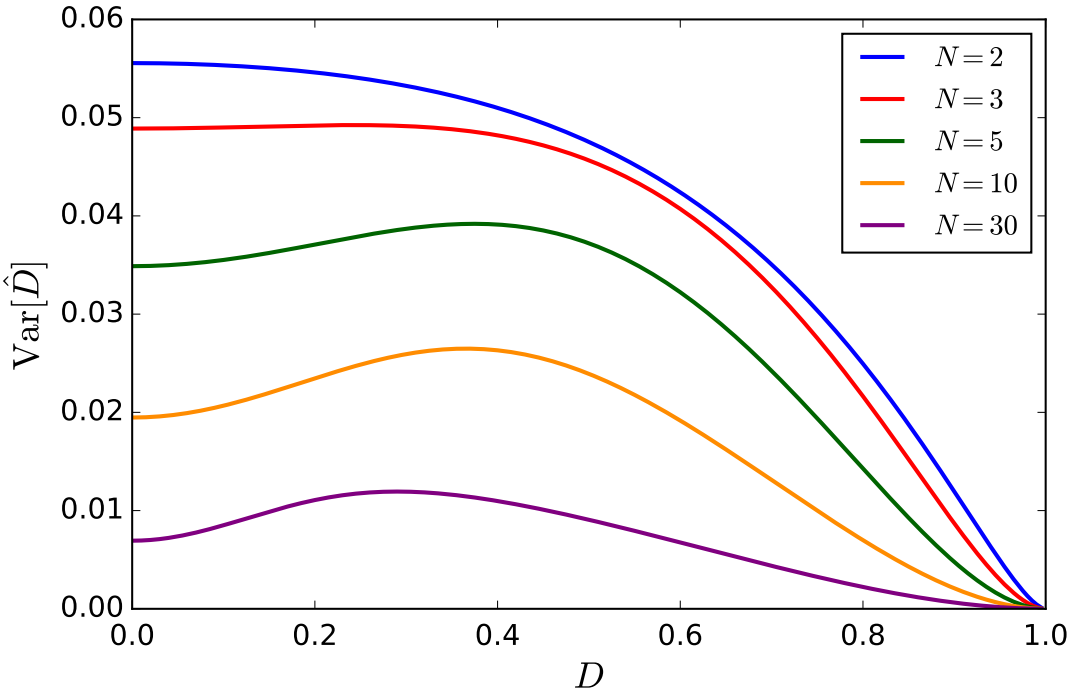
Speckle noise is only one of many possible sources of decorrelation or loss of coherence. For repeat-pass systems, the total coherence (or total correlation coefficient) of a system can be modelled as a product of multiple coherence factors [Rignot and van Zyl 1993][Santoro et al. 2007]. In the case of repeat-pass SAS, the total coherence γ can be expressed as [Bellec et al. 2005][Barclay 2006]:

$$\gamma = \gamma_n \gamma_m \gamma_b \gamma_p \gamma_t, \tag{2.35}$$

where γ_n models loss of SNR due to acoustic noise, γ_m describes the drop in coherence resulting from image misalignment, γ_b represents spatial baseline decorrelation, γ_p is associated with processing noise, and γ_t describes the temporal decorrelation of the scene. In general, these factors are not constant throughout a scene. In the context of change detection, the ideal scenario is where all coherence factors other than γ_t are close to unity, so that the sample coherence can be used to estimate the effective values of γ_t



(a) Mean estimate of the degree of coherence. The dotted line represents the true degree of coherence.



(b) Variance of the estimator.

Figure 2.4: Mean and variance of the estimated degree of coherence for different values of N, the number of looks.

throughout an image, indicating change in the scene between sonar runs. It is worth noting that a pair of speckle images with a degree of coherence of γ between them has a degree of coherence of γ^2 between their intensity images. The following subsections give a brief account of the nature and impact of each source of decorrelation.

2.6.1 Acoustic noise

Acoustic noise in sonar can come from a range of sources, such as: low frequency noise from propeller cavitation [Ross 1976][McKenna et al. 2012], ambient noise from other ships [Andrew et al. 2002][McDonald et al. 2006], flow noise (caused by the movement of water around the transducers), breaking waves [Wilson Jr et al. 1985], precipitation (rain [Nystuen 1986] or snow/hail [Scrimger et al. 1987] on the ocean surface), and marine life (snapping shrimp [Au and Banks 1998], echolocating animals, fish).

Acoustic noise is generally additive and assumed to be Gaussian, where the effect of the noise on coherence loss depends on the SNR [Zebker and Villasenor 1992][Just and Bamler 1994]. If the noise over two sonar runs is independent, with equal SNR (after image formation), then the acoustic coherence factor is

$$\gamma_n = \frac{1}{1 + \mathrm{SNR}^{-1}} = \frac{\mathrm{SNR}}{\mathrm{SNR} + 1}. \tag{2.36}$$

This relationship is shown in Figure 2.5, where, for example, an SNR of 20 dB results in a coherence of $\gamma_n \approx 0.99$. Acoustic noise is an important noise factor as it defines an upper limit on the coherence that cannot be overcome with post-processing. In the case of InSAS with multiple-receiver arrays, the acoustic noise may be highly correlated between transducers, whereas the same is not true for interferometric SAR as the additive noise predominantly comes from the electronics of each individual receiver.

2.6.2 Footprint shift

A sonar ping has a spatial width that is the speed of sound multiplied by the effective pulse duration. This pulse width corresponds to a footprint width when projected onto the seafloor at a given range. Whether the system is interferometric, multiple-receiver, or repeat-pass, the hydrophones inevitably have an offset from each other such that two footprints may largely overlap but are not exactly aligned. In the case of an interferometric sonar with an array of hydrophones, the offsets between the hydrophones result in footprint shifts. However, these footprint shifts can be corrected via interpolated delays using knowledge of the relative positioning of the tranducers, the sonar altitude, and the local geometry of the seabed.

Footprint shift is equivalent to the misregistration between two reconstructed images, as the corresponding pixels in each image do not represent the same areas of the

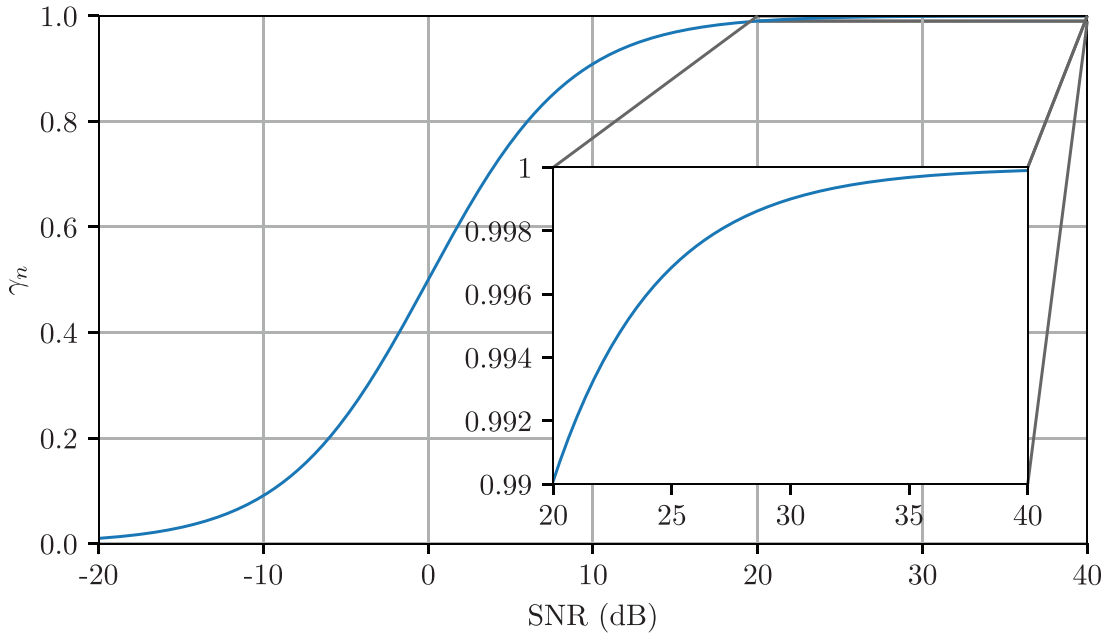


Figure 2.5: Decorrelation due to additive Gaussian noise as a function of the SNR, as given by (2.36) [Bonnett 2017].

scene. For a rectangular aperture, a misregistration by an offset α, as a fraction of the resolution size, results in a coherence of [Just and Bamler 1994]:

$$\gamma_m = \operatorname{sinc} \alpha. \tag{2.37}$$

This relationship is plotted in Figure 2.6. A footprint shift or misregistration by greater than one resolution cell results in a total loss of coherence. A common rule of thumb is that registration to within a tenth (or one eighth [Just and Bamler 1994]) of a resolution cell accuracy is required to minimise the loss in coherence from misregistration compared to other sources of noise [Persons et al. 2002][Scheiber and Moreira 2000]. This is particularly relevant for InSAS and repeat-pass interferometry [Sæbø et al. 2011][Dillon and Myers 2014b], where accurate alignment is critical for obtaining a high quality interferogram. Decorrelation due to a misregistration also carries over to subsequent processing such as coherence estimation or change detection; the quality of the results is greatly dependent on the alignment accuracy.

2.6.3 Baseline decorrelation

When the scene is imaged from different positions, the deterministic speckle pattern changes according to the baseline, which is the displacement between the transducers for InSAS or the displacement between the sonar tracks for repeat-pass imaging. Since speckle is the superposition of echoes from multiple independent scatterers within a resolution cell with random phases, a change in position is also a change in distance

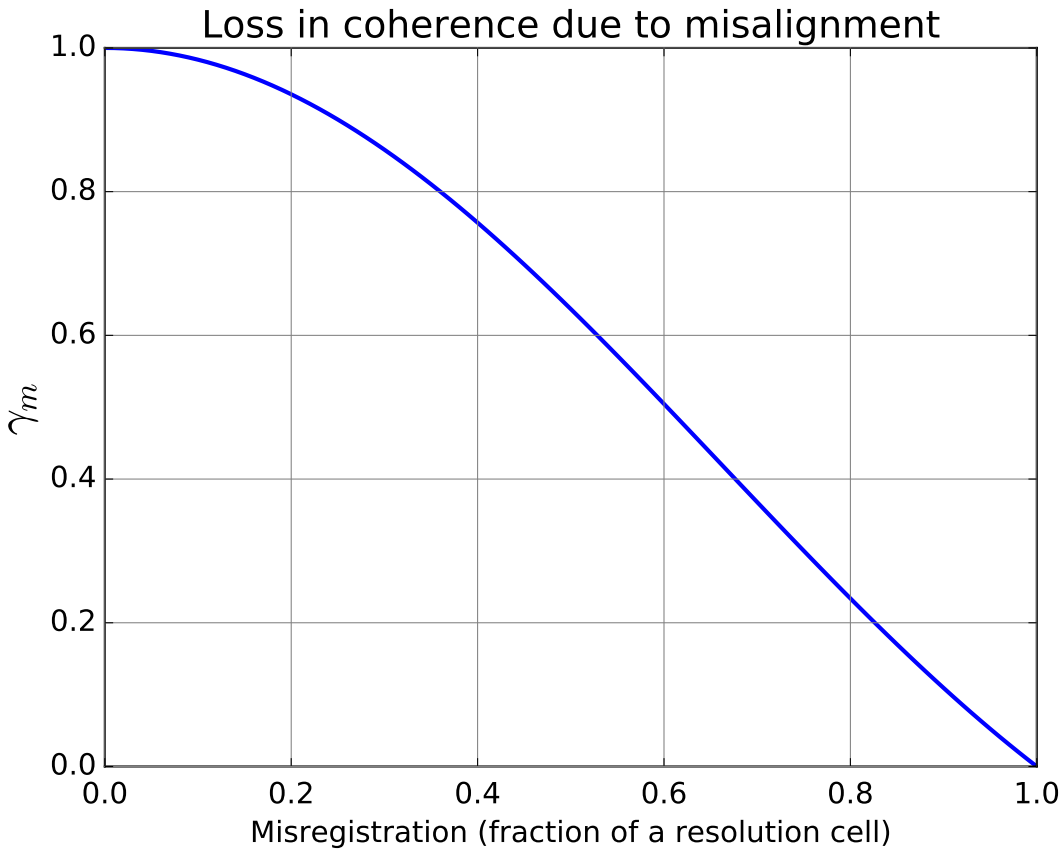


Figure 2.6: Decorrelation as a result of footprint shift or misalignment, given by (2.37). A shift of one resolution cell (or greater) results in total loss of coherence.

and aspect, and thus a change in the amplitude and phase of the signals. Thus, there is a loss of coherence that depends on the baseline [Jin and Tang 1996]. Generally, a speckle pattern remains highly correlated when the baseline is short and becomes less correlated as the angular separation increases. However, for interferometric systems, there is a tradeoff between the level of baseline decorrelation and the accuracy of height estimation [Dillon and Myers 2014a], for which an optimal baseline (for a given SNR) achieves a reasonable balance [Li and Goldstein 1990].

The geometry for baseline decorrelation and the expected loss of coherence with baseline can be found in [Li and Goldstein 1990][Barclay 2006][Zebker and Villasenor 1992]. Baseline decorrelation cannot be corrected with post-processing and is a more significant source of decorrelation for repeat-pass systems than for InSAS.

2.6.4 Processing noise

Synthetic aperture reconstruction consists of multiple stages that can each contribute noise to the final result. Three factors of noise are the: use of finite precision arithmetic; noise introduced by the interpolation methods used; and the appearance of grating lobes

due to synthetic aperture reconstruction with a finite along-track sampling rate. Each these sources are discussed in [Barclay 2006].

2.6.5 Temporal decorrelation

Temporal decorrelation is the loss of coherence due to changes in the scene between successive runs. In the ideal case for change detection, regions without any change yield a coherence of one, while regions with change have a low coherence. However, temporal decorrelation cannot be measured in isolation unless all the other coherence factors are known to be close to one. Suppose the highest sample coherence (for a suitable window size) in a scene is 0.5, which is taken to be the highest total coherence appearing in the scene. Given the high variance and bias of the sample coherence estimator (see Section 2.5), even regions of the scene that have completely changed are difficult to reliably identify; it is ambiguous whether regions with low sample coherence are caused by a drop in other coherence factors, temporal change, or pure chance.

Sediment transport is an inherent source of temporal decorrelation and depends on the sediment and environment. In shallow coastal areas, the seafloor sediment is more prone to movement due to wave action. Jackson et al. [1996] showed the temporal decorrelation of a sandy scene near Panama City (Florida, USA) to be two orders of magnitude more rapid than the decorrelation of a silty scene in Eckernförde Bay, Germany. Lyons and Brown [2013] used a rail-mounted SAS system to analyse the temporal decorrelation of a sandy scene with different sonar frequencies, showing that coherence decays more quickly for higher frequencies. The scene had notable bioturbation activity, supporting the observation that under some conditions, total decorrelation can occur within hours depending on the sonar frequency. Jackson et al. [2009] proposed a statistical model for the decorrelation of a scene over time based on the diffusion equation, showing reasonable approximations to data from both stereo-pair photographs and sonar images.

The allowable period between repeat passes for change detection to be feasible for tracking subtle changes in the scene is largely dependent on the physical and biological environment. Some sonar studies have observed changes within minutes [Roderick et al. 1984][Sæbø et al. 2011][Dillon and Myers 2014b], while incoherent change detection has also been performed over an interval longer than a year [Midtgaard 2013][Hansen et al. 2014]. Since coherent processing is far more sensitive to subtle changes in the scene, it is also likely to require significantly shorter periods (e.g., up to days) between repeat passes.

Chapter 3

Finding correspondences via feature matching

Image registration is the process of overlaying two or more images of the same scene, where the images may be taken at different times, with different sensors, from different viewpoints, and so on. Registration is a broad topic with many different specific applications; overviews of these are provided by [Brown 1992] and [Zitova and Flusser 2003]. Although image registration can be performed with multiple images, for the purposes of demonstration this thesis deals exclusively with the case of registering pairs of images.

There are multiple classes of registration methods, with correlation-based methods being the most well established. The standard correlation approach accounts for transformations between images via scale, rotation, and translation. However, only a small range of variation is allowed in these parameters, as computation costs become unmanageable when increasing the parameter space or introducing other transformations [Brown 1992]. In cases where the "pose" of a camera or sensor can vary significantly across images, correlation is often infeasible. Another well-known approach is to perform image registration using point correspondences. A point correspondence is a pair of points (one in each image) that correspond to the same geometric aspect in the real world. Given a sparse set of point correspondences, a mapping between the two images can be estimated using a suitable geometric model. The point mappings used for image registration, known as control points, can be obtained either manually or automatically. A manual method may consist of a human identifying distinctive landmarks that appear in both images and estimating their location in each image (a process that can be computer-assisted in some cases). This can be tedious and prone to inaccuracies. Automated methods have used a variety of candidates for control points, such as statistically distinctive image patches, closed-boundary regions, line intersections, and distinctive points of curvature on contour lines, along with various techniques for localisation [Goshtasby 1988a]. After the publication of the Scale Invariant Feature Transform (SIFT) [Lowe 1999], the use of local image features (especially scale-invariant blobs) became the main candidate for optical images due to their robustness and general applicability.

This chapter provides a background on local features and how they are used to

find point correspondences that can be used for image registration. A focus is given on the popular combination of SIFT and RANSAC (Random Sample Consensus) for estimating geometric relationships between two optical images, serving as an example on which the proposed feature-matching pipeline in Chapter 5 is based. SIFT is the main feature algorithm appearing in this thesis, although some results with SURF (Speeded Up Robust Features) are also presented.

Section 3.1 introduces the concept of features, feature detection and description, and the SIFT algorithms. Section 3.2 describes how a set of point correspondences can be computed from the output of feature detection and description. Section 3.3 explains the need for robust matching and presents the RANSAC algorithm. Section 3.4 briefly describes the process of geometric estimation (such as homography estimation) using point correspondences. Performance metrics for evaluating the effectiveness of algorithms for each processing stage are considered in Section 3.5. SURF and other works are briefly described in Section 3.6. A brief account of RANSAC variants and alternative methods is given in Section 3.7.

3.1 Feature detection/description and SIFT

Local image features have been used successfully in a wide range of computer vision applications such as pose estimation, scene recognition, and camera calibration [Tuytelaars and Mikolajczyk 2008]. A local feature is an image pattern that is distinct from its neighbourhood region. Features are used to identify interesting or distinctive parts of an image that are likely to be found in another image of the same scene or objects, for the purpose of finding image correspondences. Although there are many geometric entities that can be distinctive, such as edges and small image patches, the term "feature" often refers to interest points due to their sheer success for producing point correspondences in general applications. Interest points are usually identified by corner detection or blob detection, with several notable examples such as the Harris corner detector [Harris and Stephens 1988], the Hessian corner detector, and blob detectors based on the Laplacian of Gaussian. Although corner/edge detection and image features are usually applied to single-channel (greyscale) images, with conversion performed where necessary, there are ongoing efforts to develop techniques for working with colour images [Gevers et al. 2012].

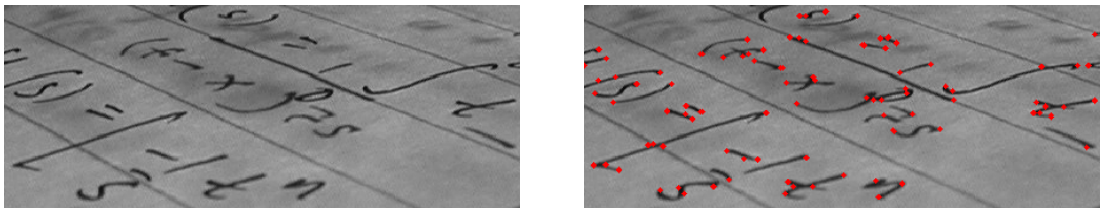
Feature detection determines and locates the position of suitable features in an image. A *feature detector* is an algorithm (usually based on a mathematical expression) that detects features. *Feature description* is the process of describing what a feature is, compressing that information using a data representation known as the *feature descriptor*. An instance of such data for a given feature is called a *feature vector*. However, the term "feature descriptor" in literature can refer to any of these three concepts: the algorithm used to perform feature description; the data representation format; or a fea-

ture vector. *Feature matching* uses sets of feature vectors to determine which features (across different images) correspond to the same real-world visual element [Tuytelaars and Mikolajczyk 2008].

There are many desirable properties of a feature detector/descriptor such as robustness, density and repeatability of detected features (which affects number of resulting feature matches), distinctiveness of the descriptor (which affects feature matching accuracy), computational efficiency, and localisation error [Tuytelaars and Mikolajczyk 2008]. For image registration applications, localisation accuracy has a heightened importance [Suri et al. 2009].

A corner can be defined as the intersection of two edges, or alternatively, a point at which there are two different dominant edge directions. An edge is a local maximum of the image gradient in a single direction. The earliest interest point detectors were simple corner detectors, which provide a structured measure of local gradient around a point, equivalent to correlation with a 2D kernel. Locations at which the response is highly positive or highly negative indicate corners with high contrast. Many of these corner detectors can be used to find sub-pixel corner locations at local maxima in the kernel response. Although these simple corner detectors are relatively efficient to compute, they are not robust to common deformations that occur between images of the same objects or scene. An example of corner detection is shown in Figure 3.1, where many of the detected corners are at ends of lines and intersections of lines of the handwriting.

After Lindeberg [1993][1994] established the supporting theory for scale-invariant blob detection, these detectors were shown to be capable of high repeatability as well as robustness to a wide range of local transformations. The SIFT algorithm [Lowe 1999][Lowe 2004] was the first major breakthrough and proposed what is now the canonical pipeline of feature detection, description, and matching used in numerous applications.



(a) (b)

Figure 3.1: Example of corner detection. (a) Source image. (b) Image with red dots to indicate the locations of detected corners.
"Output of a typical corner detection algorithm", 2006, via Wikimedia Commons (public domain).

SIFT uses the concept of scale space, which refers to the level of Gaussian blur

applied to an image $f(x, y)$. A 2D Gaussian convolution kernel is parameterised by a standard deviation, σ, and is given by

$$g(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \mathrm{e}^{-\frac{\left(x^2+y^2\right)}{2\sigma^2}}. \tag{3.1}$$

The smoothed image is $g(x, y, \sigma) * f(x, y)$, where $*$ denotes convolution in x and y. Under a set of chosen assumptions that form the scale-space axioms [Duits et al. 2004], the Gaussian kernel is the most well-known blurring filter that is guaranteed to reduce existing image structures without creating any new structures [Lindeberg 2011]. While blurring reduces available information, it can also make an image structure more mathematically or geometrically distinctive. Specifically, image structures with spatial extent significantly smaller than the scale parameter σ are greatly suppressed. Figure 3.2 shows an image at four different scales, i.e., levels of Gaussian smoothing. The scale-space representation is the family of blurred images $L(x, y, \sigma)$ such that

$$L(x, y, \sigma) = g(x, y, \sigma) * f(x, y), \quad \sigma \geq 0. \tag{3.2}$$

One of the earliest and most common blob detectors is based on the Laplacian of Gaussian (LoG), where an input image is blurred (typically via convolution) using a Gaussian kernel of a chosen scale, and the Laplacian operator is applied to the result, yielding a 2D grid response with strong positive values where there are dark blobs of a fixed radius and strong negative values where there are bright blobs of the same size. Figure 3.3 shows the shapes of the Gaussian filter, the partial derivative of the Gaussian, and the Laplacian of Gaussian, where the Laplacian of a function f is defined as:

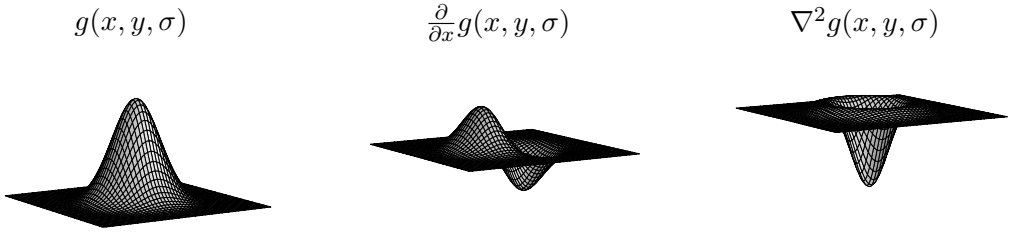
$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}. \tag{3.3}$$

Increasing or decreasing the scale of the Gaussian kernel corresponds to detecting larger or smaller blobs respectively. However, the response of the Laplacian of Gaussian also depends on the scale. Since an image likely consists of blobs of varying sizes, a solution to detecting these blobs is to normalise the LoG response according to scale. The result of this is the scale-normalised Laplacian operator,

$$\nabla^2_{\mathrm{norm}} f = \sigma^2 \left(\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \right), \tag{3.4}$$

where the scale-normalised Laplacian of Gaussian filter, $\sigma^2 \nabla^2 g(x, y, \sigma)$, can be used to detect local extrema in both space and scale. An example of multi-scale blob detection is shown in Figure 3.4, where both the position and extent of detected blobs according to scale are represented by red circles. A powerful property of using the scale-normalised

Figure 3.2: An optical image shown at four different scales [Lindeberg 2013a]. For each value of s, the original image has been smoothed by a Gaussian kernel with $\sigma = \sqrt{s}$, where image structures smaller than $\sqrt{s}$ have largely been suppressed.
Reprinted from *Advances in Imaging and Electron Physics*, volume 178, Tony Lindeberg, Generalized Axiomatic Scale-Space Theory, Copyright 2013, with permission from Elsevier.



(a) Gaussian kernel (b) Derivative of Gaussian (c) Laplacian of Gaussian

Figure 3.3: 2D Gaussian filter, its derivative, and the Laplacian of Gaussian.

Laplacian of Gaussian is that the scale-space extrema in an image are theoretically guaranteed to also appear as the corresponding extrema in a scaled, translated, and/or rotated version of the same image. The mathematical theory behind this is detailed in [Lindeberg 1998]. Further analysis of scale-space interest point detectors is given

in [Lindeberg 2013b]. In place of the Laplacian operator, another popular option is to use the determinant of the Hessian (DoH) on the Gaussian blurred image. The Hessian matrix $\mathcal{H}(x, y, \sigma)$ at a specific image location and scale is given by

$$\mathcal{H}(x, y, \sigma) = \begin{bmatrix} L_{xx}(x, y, \sigma) & L_{xy}(x, y, \sigma) \\ L_{xy}(x, y, \sigma) & L_{yy}(x, y, \sigma) \end{bmatrix}, \tag{3.5}$$

where L_{xx}, L_{xy}, and L_{yy} are partial derivatives of $L(x, y, \sigma)$, and the determinant is

$$\det \mathcal{H}(x, y, \sigma) = L_{xx}(x, y, \sigma)L_{yy}(x, y, \sigma) - {L_{xy}}^2(x, y, \sigma). \tag{3.6}$$

This can also be normalised according to scale, yielding the scale-normalised determinant of the Hessian, which shares the same scale-invariant properties as the scale-normalised LoG and also detects saddle points. According to Lindeberg [2015], this Hessian blob detector performs better than the LoG blob detector, especially under non-Euclidean affine transformations. While the LoG gives the best notion of scale, it triggers on edges more frequently (which is undesirable for a blob detector). The SURF algorithm uses a simplified form of the scale-normalised DoH for feature detection [Bay et al. 2006].

All scale-invariant feature detectors have the same basic property of invariance to scaling, translation, and rotation. However, what is referred to as invariance is more accurately called *partial invariance*; in practice, invariance to a certain transformation or deformation can only be achieved to a certain degree. Among different scale-invariant feature detectors and descriptors, there are observable differences between the degrees to which they are partially invariant to scale, rotation, translation, as well as other deformations such as illumination changes, image noise, and perspective transformations [Mikolajczyk and Schmid 2004][Krig 2016]. Many common blob detectors can also be adapted to be invariant to affine transformations. Affine invariance can be achieved by iterative warping of the smoothing kernel to match the image structure around a detected blob, resulting in improved feature description [Lindeberg and Grarding 1997][Baumberg 2000][Mikolajczyk and Schmid 2004]. Affine versions of detectors exist for the LoG, Difference of Gaussian (DoG), and DoH filters [Mikolajczyk and Schmid 2001], the Harris corner detector [Mikolajczyk and Schmid 2002], and other detectors.

The SIFT detector finds extrema within scale space using a DoG formulation, which can be shown to approximate the scale-normalised LoG [Lindeberg 2015][Lowe 2004]. A DoG image $D(x, y, \sigma)$ is given by

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma), \tag{3.7}$$

such that a DoG image at a given scale can be computed as the difference of two Gaussian-blurred images with nearby scales related by a multiplicative factor of k

Figure 3.4: Multi-scale blob detection on a greyscale image of a butterfly. Red circles are used to indicate the location and scale of the features.
"Monarch butterfly on purple coneflower" by Jim Hudgins, 2017, via Flickr (modified under CC BY 2.0).

between the scales. DoG implementations typically use a scale-space pyramid representation for efficiency reasons. With SIFT, the scale space is divided into octaves, where the ratio between octaves is a factor of two in total blur applied. (Cascading of multiple Gaussian blurs is equivalent to performing a single Gaussian blur with $\sigma^2 = \sum_i \sigma_i^2$.) The image at the base of the pyramid is pre-blurred with a kernel of 1.6. (More specifically, the input image is upsampled by a factor of two using linear interpolation. Assuming the input image is unaliased ($\sigma \geq 0.5$, adding a blur with $\sigma \geq 0.6$ results in the doubled image having at least $\sigma \geq 1.6$, which was found to be optimal by Lowe [2004]. This doubling of image size significantly increases the number of detected blobs, as a higher number of features are found at the lower levels of the pyramid.) Each octave is divided into an integer number of scales, s, such that the constant factor is $k = 2^{1/s}$. Each octave requires $s + 3$ blurred images from which to compute $s + 2$ DoG images (by subtraction from adjacent scales). Thus, extrema detection can be performed at s scales per octave. Maxima and minima within the scale space (approximated by the DoG levels) are simply pixel values that are the most

extreme within their 3×3 neighbourhood of 26 adjacent values. By default, SIFT uses $s = 3$ sample DoG levels for each octave in scale space, which requires five DoG layers computed from six scales of progressive blur per octave. The third-most blurred image in an octave, which has twice the level of blur of the octave's base image, is down-sampled by a factor of two in order to form the base image of the next higher octave. (This method of subsampling does not introduce aliasing because of the decrease in image structure due to blurring.) A total of three octaves are sampled. This scale-space pyramid representation allows detection in the higher octaves at a minor overhead cost. Figure 3.5 shows the first two octaves of the image pyramid in the case of two samples per octave ($s = 2$, $k = \sqrt{2}$), depicting the sampling of the scale space, calculation of $D(x, y, \sigma)$ from adjacent Gaussian-blurred images, extrema detection, relationship between octaves, and progressive down-sampling.

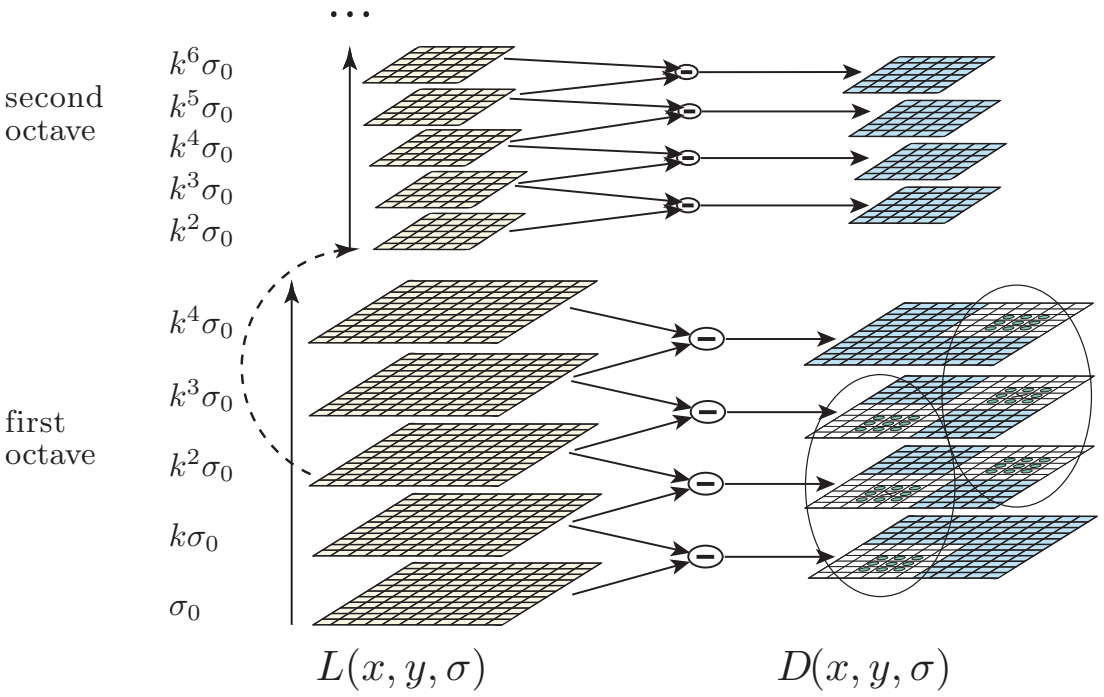
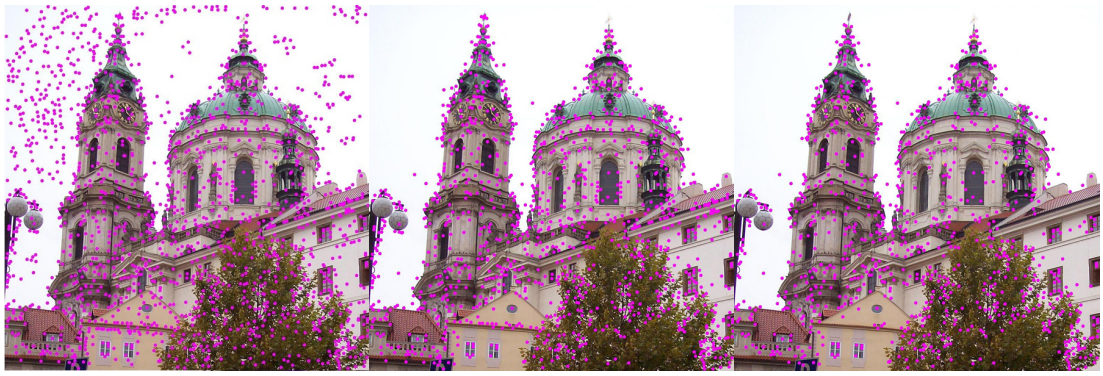


Figure 3.5: The first two octaves of the image pyramid, with two samples per octave. Each octave has five Gaussian-blurred images at progressively increasing scales, from which four DoG images are computed. Extrema detection is performed on the middle layers. The third most blurred image in the first octave is downsampled to create the base image of the second octave.
Adapted from [Younes et al. 2012]. Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature. *International Journal of Computer Vision*. Distinctive Image Features from Scale-Invariant Keypoints. David G. Lowe. © 2004.

Extrema detection produces many features (known as keypoints in the SIFT algorithm), some of which are unstable or inaccurate. The next step is the refinement of keypoint locations using interpolation of the DoG images. Interpolation is performed

using a quadratic Taylor expansion of the DoG function based on the local derivatives (gradients) and solving for a stationary point. If the estimated extremum is closer to a different pixel sample point, the same estimation procedure is repeated using that sample point instead.

After the computation of keypoint locations, keypoints in regions of low contrast are discarded. Specifically, the absolute value of the interpolated DoG function at a location is compared to a user-specified threshold known as the contrast threshold. Both edges and corners have a high response in scale space, but edges tend to be more ambiguous, have unreliable locations, and be more prone to the effects of noise. (Edges are defined as having a significantly higher principle curvature along one direction than another.) Keypoints that lie on edges are eliminated based on the eigenvalues of the second-order Hessian matrix, whose values are proportional to the curvatures of the DoG function. The effects of constrast thresholding and edge removal can be seen in Figure 3.6.



(a) (b) (c)

Figure 3.6: Detected SIFT keypoints. (a) Maxima and minima of the DoG images (scales not indicated). (b) Remaining keypoints after thresholding with a minimum contrast parameter. (c) Final keypoints after edge removal using a threshold on the ratio of principal curvatures (see edges of the buildings).
"SIFT keypoints filtering" by Lukas Mach, 2008, via Wikimedia Commons (modified under CC BY 3.0).

The final stage of SIFT keypoint detection is the assignment of orientations, where each detected keypoint location produces one or more keypoints with distinct orientations. Keypoint orientations are used to achieve rotational invariance, as feature descriptors can be used to represent local image information relative to detected orientations that are repeatable in theory. Computation of orientation is performed using the Gaussian blurred image at the scale of each keypoint, which ensures invariance to scale. The gradient magnitude and orientation are computed using the pixel differences of the smoothed image. The magnitudes and orientations are sampled at multiple points within a region around the keypoint to form an orientation histogram covering

360 degrees with 36 bins. The contribution of each sample to its orientation bin is the gradient magnitude multiplied by a Gaussian weighting extending from the keypoint location with 1.5 times the scale of the keypoint. The maximum peak in the orientation histogram is determined, and all orientations with greater than 80% of that peak value become the orientations of separate keypoints at the same location. Finally, quadratic interpolation of the peak is used to refine the orientation of each keypoint.

After feature detection, each keypoint has a location, scale, and orientation. The next stage is feature description, which represents the local image data around the keypoint in a scale-invariant manner such that these image structures can be identified as repeated keypoints from other images by comparing these representations for simlarity. Instead of sampling image intensities, which is remniscent of correlation-based matching, SIFT uses an approach inspired by the behavior of complex neurons in the mammalian primary visual cortex, which responds to gradients at a certain frequency and orientation. As with orientation assignment, feature description is performed based on the blurred image with the closest scale to the keypoint's scale. The SIFT descriptor is a histogram of gradient magnitudes and orientations sampled from multiple subregions around the keypoint (see Figure 3.7). Firstly, a 4 by 4 window is formed around the keypoint, rotated according to the keypoint orientation. For each window subregion, a gradient histogram of eight orientation bins is computed. Each gradient histogram is formed by summing the contributions from a grid of 4 pixels by 4 pixels, where gradient magnitudes are weighted by a Gaussian window with σ equal to half of the scale of the keypoint such that samples further from the keypoint are less influential. It is crucial to ensure the histograms are stable; boundary effects that would be caused by subtle changes in the image, where samples contribute to alternative orientation bins or shift from one subregion to another, should be minimised. This is achieved by distributing each gradient sample over eight bins in total using trilinear interpolation; the two nearest orientation bins for each histogram of the four spatially closest subregions. The weightings of these contributions decrease with distance in either dimension. SIFT uses a 128-element vector to represent the $4\times4\times8$ total orientation bins for each keypoint.

Lastly, the feature vector is normalised to unit length in order to reduce the effects of changes in illumination. Normalisation provides invariance to affine changes in illumination such as scaling by a constant factor and uniform changes in brightness, but is not sufficient to handle saturation and non-linear illumination changes. Since illumination effects are often sensitive to angle, all magnitudes in the unit vector are thresholded to 0.2 and the thresholded vector is normalised again.

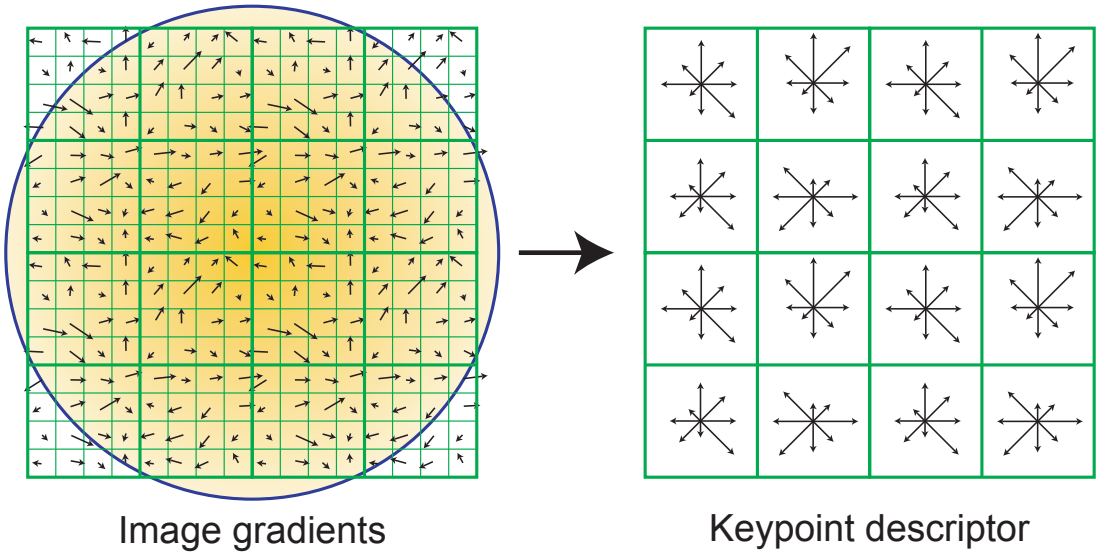


Figure 3.7: The 128-element SIFT keypoint descriptor is structured as a 4×4 descriptor array of histograms with eight orientation bins each, as shown on the right. The length of each arrow corresponds to the sum of gradient magnitudes for that orientation bin within a nearby region. Gradient magnitudes and orientations are sampled according to a 16×16 grid (shown on the left), with each sample contributing to the two nearest orientation bins across each of the four nearest subregions, additionally weighted by a Gaussian window (indicated as a circle for simplicity).
Adapted from [Lowe 2004]. Adapted and reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature. *International Journal of Computer Vision*. Distinctive Image Features from Scale-Invariant Keypoints. David G. Lowe. © 2004.

3.2 Feature matching

The SIFT descriptor is a floating point feature descriptor, as opposed to a binary feature descriptor. Floating point feature descriptors are compared for similarity using the Euclidean distance between them. Feature matching across two images finds pairs of features that have (ideally) the smallest distance of any other pair. The most basic method of feature matching is to perform a brute force search, where every feature vector in a second image is compared to each feature vector in the first image, and the feature with the lowest descriptor distance is deemed a match to the feature from the first image. Brute force matching results in a tentative feature match for each feature in the reference image. In practice, many of these matches are likely to be incorrect matches due to the repeatability of features being less than 100% and various other sources of differences between the images. One method for removing likely poor or ambiguous matches is Lowe's ratio test, where the distance between two matched descriptors (one being from the reference image) must be within a chosen ratio of the distance of its second nearest descriptor. Lowe's ratio test can eliminate a significant proportion of outlier matches while discarding relatively few inlier matches. Lowe used

a ratio of 0.8.

Applying Lowe's ratio test requires solving the 2-nearest neighbours problem. Since brute force has an costly overall running time (quadratic in the number of features), a popular alternative is to use the Fast Library for Approximate Nearest Neighbours (FLANN) library, which uses randomised kd-trees to perform an approximate nearest neighbour search [Muja and Lowe 2009][Muja and Lowe 2014]. There are no known algorithms that are faster than brute force while also giving the exact closest matches [Muja and Lowe 2009].

In the case of SIFT features, it may be appropriate to remove redundant correspondences prior to estimation, since equivalent matching feature locations are not statistically independent [Rabin et al. 2010].

3.3 Removing outliers using RANSAC

After obtaining a set of tentative feature matches, regardless of whether Lowe's ratio test is performed, outlier feature matches still typically remain. While reasonable image registration estimates may be attainable based on models of the imaging geometry in the presence of noise, these models typically cannot cope with outliers. Therefore, the goal of consensus-based outlier removal is to find as large a set as possible consisting only of inliers. The canonical method for outlier removal of feature matches is RANSAC [Fischler and Bolles 1981]. Although RANSAC is a robust estimation algorithm that can be applied to other problems, a description specific to the context of feature matching is given here.

An iteration of RANSAC consists of the following steps:

1. Randomly select a set of m sample matches from the whole set of feature matches and compute an estimate of the geometric model, where m is the minimum number of correspondences required to form an estimate.

2. Determine the number of matches that are consistent with the predicted model, where a pair of matched points is consistent if their distance from the modelled relationship is within a chosen error threshold. This requires a definition of the reprojection error.

RANSAC typically runs for a fixed number of iterations, where the final model is re-estimated from the largest consistent set that was encountered. The matches that are consistent with the final solution are deemed to be correct feature matches. (There are several variations of RANSAC; in the original, the algorithm terminates as soon as a sufficient number of consistent data points is found.) RANSAC is a nondeterministic method and is not guaranteed to find a correct set of inliers. If the proportion of inliers in the tentative set of matches is p, then the probability of choosing a random set of m

inliers is approximately p^m. After N iterations, the probability of encountering at least one set consisting of only inliers is $1 - (1 - p^m)^N$. Thus, given a desired confidence ϵ of finding a correct solution, the minimum number of iterations required is:

$$N \geq \frac{\log 1 - \epsilon}{\log 1 - p^m}. \tag{3.8}$$

.

3.4 Geometric estimation from point correspondences

In the field of computer vision, RANSAC is the typical choice of algorithm for obtaining a set of inlier matches given tentative matches from feature matching. Although the final estimate from RANSAC is computed from and is consistent with all the inlier data, the estimate can usually be further refined. Common definitions of the reprojection error function and the nature of RANSAC serve to account for as many inliers as possible [Hartley and Zisserman 2003], a goal that is distinct from minimising the effect of noise in trying to find an accurate estimate. Although it may be sensible to perform refinement on the estimate from RANSAC, this is not always done in computer vision applications due to the relative stability of estimation with optical camera models.

The minimum number of correspondences required to estimate a direct relationship between two images depends on the geometry, as well as the property being estimated. Two-view geometry is the relative geometry of two different perspective views of the same 3D scene and imposes three geometric constraints. The pin-hole camera model is assumed. The *epipolar geometry* fully describes the geometric relationship between two cameras, and can be estimated from seven or eight point correspondences [Hartley and Zisserman 2003]. Two images from different perspectives of the same planar surface in a scene are related by a planar homography, which can be solved from four correspondences. This is equivalent to calculating the *projective warp* that transforms one 2D surface onto another [Marburg 2015]. An affine transformation can be calculated from three correspondences [Goshtasby 2005].

3.5 Performance metrics

It can be difficult to objectively characterise and compare the performance of different feature detectors. A basic aspect of performance is the number of features found by a detector. Too many detected features can correspond to unnecessary overhead, whereas too few features can lead to unstable estimation or total implausibility of geometric estimation. Many feature detectors have parameters that affect the sensitivity of detection. For example, SIFT finds more features when the contrast threshold parameter is lowered. However, the default parameters of feature detectors are usually suitable for

general purposes, and adjusting these for the sake of acquiring more features usually results in a more significant loss of quality in other metrics. Another important metric is the repeatability of features, i.e., the likelihood or proportion of the same features being detected and matched across different images. This depends on the detector, and to some extent, the descriptor. The repeatability varies widely according to image content, but the relative performance of different feature detectors can be concluded from comparative studies performed using the same image datasets. While estimates for feature repeatability cannot be generalised, the raw number of detected features is of no value unless the features are repeated, otherwise unmatched features represent either features that do not appear in both images or detections purely due to noise. The overall computational costs of performing feature matching is also a significant aspect. For example, real-time performance of feature matching generally cannot be achieved using SIFT and SURF. (Detectors and descriptors considered suitable for real-time applications include FAST (detector) [Rosten et al. 2010], BRIEF (descriptor) [Calonder et al. 2010], FREAK (descriptor) [Alahi et al. 2012], ORB (descriptor) [Rublee et al. 2011], BRISK (descriptor) [Leutenegger et al. 2011], and STAR (detector, derived from CenSurE [Agrawal et al. 2008]). A summary and performance comparison of these can be found in [Patel et al. 2014].) When running on smaller embedded systems, large feature vectors (such as with SIFT) can be discouraging in terms of both memory footprint and computation time of the matching stage. The primary emphasis of a feature descriptor is to offer good distinctiveness between dissimilar features. This property is difficult to measure objectively and in isolation. For example, the SIFT descriptor was designed with SIFT features in mind, so it may perform relatively worse when paired with feature detectors that are dissimilar. Although rarely discussed in literature, the sub-pixel localisation accuracy of detected features has a theoretical impact on the resulting estimations using the correspondences. In terms of feature matching output, an important measure is the ratio of inlier matches before RANSAC, since RANSAC can struggle to find a consensus set when the inlier ratio is low. Additionally, it is desirable for RANSAC to discard as few true inliers as possible among the rejected outliers.

3.6 SURF and other works

Speeded Up Robust Features (SURF) [Bay et al. 2006][Bay et al. 2008] was designed as an improved detector and descriptor over SIFT, mainly known for its faster computation. As a scale-invariant blob detector, it is conceptually similar to SIFT, although it achieves the desired properties of invariance using different mathematical computations. Being based on the Hessian matrix, the SURF detector uses an approximation to the DoH for interest point detection. Whereas SIFT uses the DoG operator to approximate the LoG, SURF approximates the second-order Gaussian derivatives using discrete box filters with appropriate size increments based on scale. This aggressive approximation

results in a small decrease in accuracy but a significant speed improvement, allowing for fast filtering regardless of filter size (or scale) by using integral images [Viola and Jones 2004]. The process of applying Gaussian blurs at successively increasing scales is thus replaced by applying box filters with increasing size. The Frobenius norm of the filters is kept constant so that they are effectively scale-normalised [Lindeberg and Bretzner 2003]. Like SIFT, the scale space is divided into octaves with a fixed number of samples per octave for the sake of extrema detection. Non-maximum suppression [Neubeck and van Gool 2006] in a 3×3×3 neighbourhood is applied, followed by interpolation of the maxima in scale space using the method proposed by Brown and Lowe [2002].

A dominant orientation is determined for each detected interest point based on Haar wavelet responses in the horizontal and vertical directions within a circular neighbourhood of radius six times the scale of the feature. The wavelet filters are also scale-dependent in size and can be applied efficiently using integral images. Each response is weighted by a Gaussian window centred on the feature location. Using a sliding orientation window on the 2D space of gradient directions and magnitudes, the sum of gradients within the orientation window is computed, with the largest resulting vector for any window being chosen as the orientation of the interest point. For feature description, a square region with width twenty times the scale is constructed around the interest point, oriented according to the dominant orientation. This region is divided into 4×4 square sub-regions. For each sub-region, Haar wavelet responses are sampled according to a 5×5 grid and weighted by a Gaussian centred on the interest point. A four-dimensional vector is calculated, consisting of the sums of responses along each rotated axis direction as well as the sums of the absolute values of these responses. Thus, the SURF descriptor has a total of 64 dimensions and is invariant to scale and a constant bias in illumination. The vector is normalised to achieve invariance to contrast. A more thorough account of the details of SURF can be found in [Oyallon and Rabin 2015].

Aside from SURF's superior computational efficiency, many adaptations of SIFT have been proposed in the form of variants: PCA-SIFT [Ke and Sukthankar 2004], which uses a more distinctive descriptor using a gradient based on principal component analysis; ASIFT [Morel and Yu 2009], a fully affine invariant descriptor (whereas SIFT is invariant to four out of six of the affine transform parameters); GSIFT [Mortensen et al. 2005], whose descriptor captures global context in order to improve feature matching under non-rigid transforms); CSIFT [Abdel-Hakim and Farag 2006], a colour invariant adaptation of SIFT; and RootSIFT [Arandjelovic and Zisserman 2012], which compares descriptor histograms more efficiently. A slight variation on Lowe's ratio test is to perform dual matching, where a pair of matched features is required to pass the ratio test symmetrically rather than only in terms of the closest two matches of the reference image [Wang et al. 2012]. Despite the volume of work following SIFT's pro-

posal, SIFT remains a popular choice of algorithm for feature detection and matching due to its robust performance. It is especially suitable for non-real-time applications and in many situations its variants are not needed.

3.7 RANSAC variants and alternatives

RANSAC (and its family of methods based on the same idea of confidence) are considered the "gold-standard" algorithm for robust pose estimation, homography calculation, and image registration in the presence of noise and outliers. RANSAC is also widely used for other parameter estimation problems, especially in the field of computer vision where data extracted from images is heavily affected by non-natural processes. RANSAC has also been used for image registration in SAS [Kim 2007] and SAR [Dellinger et al. 2015].

There are several well-known adaptations of RANSAC.

- Locally Optimised RANSAC (LO-RANSAC) [Chum et al. 2003][Lebeda et al. 2012] promises to find better inlier sets (and thus more consistent solutions) for the same number of iterations of RANSAC with minimal computational overhead. The LO-tweak is based on constructing more than the minimal set from the best-so-far inliers to form a more accurate estimate. An underlying assumption is that model estimation can be performed efficiently for non-minimal data sets.

- PROSAC [Chum and Matas 2005] preferentially chooses points that are known to be more likely inliers before converging to the uniform sampling behavior of RANSAC. This sampling strategy conserves the same solution guarantees as RANSAC but can achieve a noticeable speed-up when the quality of the data varies greatly and can be ranked by confidence.

- Preemptive RANSAC [Nistér 2005] employs a breadth-first search and keeps a best-so-far solution for when the RANSAC confidence cannot be achieved due to (real-)time constraints.

- Bail-out test for RANSAC [Capel 2005]

- R-RANSAC with SPRT [Chum and Matas 2008]

- QDEGSAC[Frahm and Pollefeys 2006] improves on the ability to find good solutions when samples are often near-degenerate.

- MLESAC [Torr and Zisserman 2000]

- NAPSAC [Myatt et al. 2002]

In some applications such as feature matching of SAR images the proportion of inliers may be somewhat low, requiring a large number of RANSAC iterations. "A contrario RANSAC" implementations are lesser known variants that can handle a low proportion of inliers ($< 10\,\%$), using an *a contrario* statistical significance model that eliminates the need for a threshold. Moisan and Stival [2004] used a contrario RANSAC to handle situations with few inliers. Sur [2010] chose an a contrario implementation in order to eliminate the need for an explicit RANSAC parameter, despite a high inlier ratio. Rabin et al. [2010] extended the idea to their novel algorithm, MAC-RANSAC, for robust matching in the case of duplicate objects in the same image.

RANSAC is not grounded in statistics, and is therefore not favoured in the field of statistics. Alternatives for robust estimation include M-estimators, LMedS, MINPRAN [Stewart 1995], robust regression methods, and other ad-hoc methods.

Chapter 4

Sonar image registration

Image registration, also referred to as coregistration, is the process of remapping multiple images onto a single coordinate system. In sonar, this means mapping a repeat-pass image onto the grid of a base image so that the reflectors are located at the same coordinates [Vallestad 2017]. With most methods, image registration consists of computing offsets between corresponding locations or features in the images, followed by warping or transforming one image onto the coordinate system of the other using these offsets. In the context of SAS, image registration is a necessary precursor to applications such as repeat-pass interferometry and change detection. In these applications, image registration must be accurate in order to produce useful results. In particular, a general rule of thumb for both SAR and SAS is that two images must be aligned to within about a tenth of a resolution cell in order for the phase information in coherent images to be salvageable for coherent processing such as computing an accurate interferogram [Just and Bamler 1994][Scheiber and Moreira 2000][Dillon and Myers 2014b][Sæbø et al. 2011].

There are several factors that make image registration a non-trivial challenge. A "perfect" coregistration of two images would be a mapping of points of a common scene to an arbitrary accuracy. The most typical and usable form of this (under the presumption of a fairly level scene without extreme height variations) would be two distortionless SAS images reconstructed at the average scene depth with identical scale and an exact known 2D alignment (translation and rotation) between the two. However, image distortion is unavoidable in all real-world sonar applications. Blurring, distortion, and other artefacts inevitably appear in reconstructed images due to a degree of unstable motion (such as drift and sway) of the sonar system while scanning the scene. These artefacts can be minimised to some extent using motion compensation techniques that use the measured motion data of the system as well as inferred information [Putney et al. 2001]. There are also artefacts caused by shadows, foreshortening, and layover [Franceschetti and Lanari 1999]. These are geometric distortions inherent to the imaging geometry and cannot be corrected except by combining data from multiple passes [Vallestad 2017].

Supposing that reconstructed imagery can be corrected for motion such that it is to scale, there is another major obstacle for image registration. If the relative motion of a sonar system along its designated track is known precisely, an image of the scene can be produced with minimal distortion. However, for two such images of the same scene to be precisely aligned requires accurate knowledge of the global positioning of each observed sonar path or accurate knowledge of the relative positioning between the two sonar paths. In practice, there is no easy way to measure the relative position of two sonar paths without having some common reference (as in global positioning). Several options for positioning systems have been tested, such as using GPS (in the case of a hull-mounted sonar) [Bonnett et al. 2013], positioning using active beacons [Smith and Kronen 1997][Willemenot et al. 2009], and using a long baseline acoustic positioning system where transponders are fixed to the seafloor near the area of interest [Pilbrow 2007]. In cases where position is not explicitly measured, rough estimates can be inferred using navigation data such as from inertial navigation systems, Doppler velocity logs (DVLs), other sensors, and sensor fusion approaches. However, in all cases, state of the art positioning and motion estimation simply cannot track positional information to a degree of accuracy sufficient for registering images accurately using such positioning/motion data alone. There is also no readily available instrumentation to provide the ground truth vertical separation between two sonar runs [Dillon and Myers 2014a]. Even a navigational accuracy to within one centimeter is not a sufficient basis for coregistration, as the required alignment accuracy in physical units may be on the order of a few millimeters [Bonnett et al. 2013]. Data-driven techniques that heavily rely on use of data from the scene (but may still incorporate navigation data) are necessary to register images to within the desired one-tenth resolution cell accuracy for coherent processing of data. Although the accuracy requirements for incoherent processing (such as incoherent change detection) are not as strict, sub-pixel registration accuracy is still required. (For example, misregistration by a whole pixel results in total loss of coherence.) However, poorly registered images may still be suitable for human interpretation.

Since satisfactory image registration requires use of the observed scene data, it follows that the measurement of the scene data must be of sufficient quality. The sonar data is generally reconstructed into images of the scene, and an image registration is calculated from the image data and not from the original sonar data. Therefore, even in the case of high quality sonar measurements, there are multiple additional factors that affect whether an image registration can be accurately estimated. For example, it is feasible to obtain sonar images that show the same region of interest on separate runs without the use of any scene data. However, if the scene of interest changes completely between runs (with no common objects or other distinctive features), then it is impossible to register them accurately (using only the relatively course navigation data) as there is no basis for determining which points in one image map to which points

in the other image to within fractions of a centimeter. (Although, arguably, there is little point in registering them in such a case in the first place.) Repeat-pass images must have a high degree of similarity in order for accurate registration to be possible. There are various measures of similarity that can be applied, with correlation (or coherence) being the best known. Decorrelation between images of the same scene can come from various sources: acoustic or thermal noise, image misregistration, baseline decorrelation (or speckle decorrelation), processing noise, and temporal decorrelation of the scene [Barclay 2006].

There are two main approaches to image registration for SAR and SAS: area-based methods and feature-based methods [Bentoutou et al. 2005], where the first task is to determine point-to-point offsets between the images. With feature-based methods, these point correspondences are sparsely distributed, whereas they may be densely located (such as in a grid) in the case of area-based methods. This chapter provides an overview of area-based registration in Section 4.1 and feature-based registration in Section 4.2. Image warping, which may be paired with either approach, is a technique for transforming one image onto the coordinate system of the other given a set of point correspondences and is covered in Section 4.3. Section 4.4 provides a brief literature review and outline of similar or related topics to registration of sonar images, including a summary of previous feature-based SAS work.

4.1 Area-based methods

Area-based methods, sometimes referred to as correlation-based methods, are the traditional approach to registration of speckle images in SAR and SAS. If two images are related by a sub-pixel 2D shift without any rotation between them or image distortion, then the shift between the images can be estimated using correlation. Specifically, the arguments at which the image correlation function is a maximum is deemed to be the offset between the two images. Since the correlation function has integer arguments, it is necessary to perform some form of interpolation in order to estimate the offset to sub-pixel accuracy. Theoretically, the ideal estimate is the 2D offset for which a sinc-interpolated version of the correlation has a global maximum. However, this result is rarely implemented due to being more difficult and computationally expensive to compute. Several approximations are available, with the most common being quadratic interpolation, oversampling using sinc-interpolation followed by linear interpolation, or oversampling followed by quadratic interpolation, with the latter giving the best result. Oversampling the correlation function by $2\times$ or $4\times$ near the suspected peak (followed by linear or quadratic interpolation) is considered to be sufficient in terms of accuracy. A comparison of quadratic interpolation and quadratic interpolation after oversampling is given by Bonnett [2017].

In terms of accuracy of peak detection, the variance of the error in estimated

across-track location is inversely proportional to the system bandwidth, although for the Cramér-Rao lower bound (CRLB) the standard deviation is inversely proportional to the bandwidth [Quazi 1981]. The error in peak detection is related to the width of the peak; when correlating two images, the shape of the peak is governed by the autocorrelation of the system impulse response, which is different in the along-track and across-track directions. For SAS, the shape of the peak in the across-track axis is a function of the bandwidth of the transmitted signal, and the along-track autocorrelation is dependent on the shape and dimensions of the transmitter and receiver [Fortune 2005].

Computing the correlation is a relatively expensive operation. Rather than calculating the cross-correlation of two whole scene images, a windowed correlation is performed, where each estimate is computed from rectangular subregions of each image. This approach implies that the two images are known to be correlated, with the objective being to determine the offset between them. In this sense, windowed correlation is roughly equivalent to performing a correlation except using the subimage of one of the images, and also taking into account the relative offset between the subimage coordinates and the original image. For a window size of $P \times Q$ and an image size of $M \times N$ (with $M \geq P$, $N \geq Q$), the naïve spatial domain implementation of windowed correlation is $\mathcal{O}(MNPQ)$ in number of multiplications required. The correlation output matrix has $M + P - 1$ rows and $N + Q - 1$ columns), where each entry nominally requires PQ multiplications over the window. The correlation can also be calculated in the Fourier domain, with a computational complexity of $\mathcal{O}(MN \log MN)$, which may be more efficient than the time domain method when the window size is large. However, using summed area tables [Crow 1984], also known as integral images [Lewis 1995][Viola and Jones 2004], the spatial domain implementation with $\mathcal{O}(MN)$ complexity can be achieved with a small reduction in numerical precision [Bonnett 2017].

Various local image distortions can arise in a synthetic aperture image, including rotational effects. Finding the optimal coregistration of a single point (or the patch around it) can be as simple as locating the peak correlation or as complicated as incorporating models of perspective transforms and searching for a set of mapping parameters over multiple iterations. Furthermore, point correspondences must be calculated throughout the image, where different regions of the image may behave differently.

4.1.1 Correlation-based methods

Bonnett [2017] proposed a registration technique with the purpose of determining the difference in heading and 2D offset between two sonar tracks using the images. This involved dividing one repeat-pass image into blocks and registering each block relative to the other image using a larger search window, thus generating a displacement field over the whole image. For each block, the displacement model was based on 2D shift and

a rotation, where the set of parameters that yielded a peak correlation was estimated using a numerical minimisation algorithm. Small rotations of an image caused sharp decreases in correlation magnitude, necessitating oversampling of the correlation function. The correlation peaks were located using sinc oversampling followed by quadratic interpolation. Finally, the set of displacement fields throughout the image was used to estimate the overall shift and rotation between the two sonar runs. Performing such a track registration can potentially lead to a higher coherence after reconstructing the image according to the new data-based track estimate. An ideal straight sonar track was assumed and detection of correlation peaks was found to be sensitive to the choice of block size.

Sæbø et al. [2011] proposed a repeat-pass interferometric SAS coregistration procedure that estimates the navigation error, adjusts the navigation data, then reprocesses the imagery. Firstly, the two repeat-pass images are projected onto the same coordinate system. Normalised cross-correlation using a 51 by 51 pixel window is used to estimate the local displacements between the images from the correlation peaks. The grid of displacements is averaged in both dimensions and a plane is fitted to model the average along-track and across-track shifts. The original navigation estimates are modified by adding the estimated shifts and the second image is regenerated. The coherence between the first image and the regenerated version of the second image is estimated using a 9 by 9 window. Next, the interferogram is averaged to produce refined shifts, and these refined shifts in the navigation are used to regenerate the second image. The coherence is re-estimated, with the across-track trend in the interferogram used to estimate a rotation. This rotation is applied to the navigation data to regenerate the second image. The coherence is re-estimated once again, and an along-track scaling (or surge error) is estimated from the interferogram to renavigate and regenerate the second image. Re-estimation is performed one final time, where the interferogram is averaged to obtain final shifts applied to the navigation data in regenerating the second image. Overall, this algorithm attempts to estimate and resolve misregistration between the original motion-compensated images due to unaccounted shift, rotation, and scaling.

It is worth noting that correlation is typically performed on complex images in sonar, whereas with radar applications the original complex images are not always available to researchers.

4.2 Feature-based registration

Feature-based registration in sonar and radar is inspired by the well-known feature matching pipeline in computer vision such as for homography estimation. Here, point correspondences are found between two images using feature matching and then robust estimation in the presence of outliers is performed using RANSAC with an estimation

model for the homography transformation parameters. The sonar imaging geometry differs from the epipolar geometry of cameras and is non-affine except in the trivial case where the sonar images are reconstructed at true ground depth prior to registration. Therefore, the transformation and estimation model used with RANSAC must be specific to sonar. In all other respects, the feature matching pipeline may be the same, however, there may also be several tweaks used that are specific to speckle imagery, such as applying speckle reduction filters, the use of feature detectors designed for speckle, modifying RANSAC, or even replacing RANSAC with other robust estimation methods.

Feature matching produces a sparse set of point correspondences, where the emphasis of the design of features is to provide reliable matches rather than accurate localisation. Although refinement of feature locations is never performed in feature matching, most feature detectors achieve sub-pixel localisation accuracy across matched features. However, correlation (see Section 2.2), as the optimal linear filter for template matching in the presence of Gaussian noise, is superior in accuracy. Correlation is utilised for its accuracy at the expense of computation time, whereas feature matching yields less accurately located, sparsely distributed matches at lower computation times. Unlike feature matching, correlation-based methods also tend to involve generating dense samples of point correspondences. Since accuracy remains of paramount importance with registration of speckle images, especially with coherent processing, the more recent role of feature-based registration of speckle images has been to provide a coarse initial estimate to be refined by subsequent correlation-based registration. This approach may provide an overall speedup since it reduces the search space for correlation methods, which can be an order of magnitude slower than feature-based registration. Whereas area-based methods are applied to complex speckle images, feature algorithms work on greyscale images. Thus, complex speckle images must be converted to sensible greyscale images through some arbitrary mapping; taking the log of the magnitude image clipped to a finite intensity range is the standard method. The dynamic range of the converted image can be chosen according to the scene content.

Feature-based methods can have an advantage over area-based methods when major illumination changes are expected or in multisensor applications [Zitova and Flusser 2003].

4.3 Image warping

Image warping is a general technique for converting a set of known displacements (usually obtained using correlation) between two images into a piecewise transformation or mapping between two images. The advantage of image warping is that it is more flexible than using a model of the imaging geometry and can account for wide variations in local artefacts and image distortions that even a piecewise geometry model cannot

capture; it is not constrained by the physics of the imaging process.

Given a set of displacement estimates distributed throughout the images, the purpose of image warping is to define the displacement at any point in the image. In the image warping approach, a 2D surface function is fitted to the displacement estimates. An example of the steps taken for image warping of radar images is as follows [Preiss and Stacy 2006]:

1. A coarse global shift in range and azimuth is estimated by correlating the whole images.

2. Based on the coarse global displacement, the images are partitioned into smaller subregions. For each subregion, the correlation is used to accurately determine the local misregistration of the subregion.

3. The local misregistrations are fitted using a pair of thin plate splines [Duchon 1977][Goshtasby 1988b], one for shifts in range and the other for azimuth.

4. Using the spline warping functions, the complex repeat-pass image is resampled onto the grid of the primary image with sinc interpolation.

5. Any dominant relative phase difference between the primary image and the resampled repeat-pass image is estimated and removed from the original repeat-pass image.

6. The local misregistrations are recomputed as in Step 2 but using the modified repeat-pass image. The peak correlations should be higher than before.

7. The warping surface is recomputed and the images are resampled to obtain a registered image pair.

Note that heading or rotation errors are far more relevant in sonar imaging due to lack of precise control over navigation compared to aerial or satellite radar. Thus, for sonar, image warping may require additional consideration of the rotation, such as using a larger search space to incorporate small rotations when estimating correlation peaks in Step 2 [Bonnett 2017] or estimation of rotation from the interferogram as an intermediate step [Sæbø et al. 2011]. Data-driven estimation of navigation error followed by image reprocessing [Sæbø et al. 2011] is more powerful as the phase effects of image distortion can be accounted for. In radar, image warping is common due to its manageable navigation errors, potential unavailability of the raw complex images, and the simplicity and efficiency of image warping. Although thin plate spline models have no particular connection to the distortions in sonar and radar imaging, they have been widely used to model non-rigid transformations in image alignment and shape matching due to its convenient properties: it produces differentiable smooth surfaces, provides closed-form solutions for both warping and parameter estimation, requires no

extra parameters, and minimises the energy of surfaces in a physical way [Bookstein 1989].

4.4 Related topics

This section provides a brief survey of background literature on related topics to feature-based registration of speckled images. Sections 4.4.1–4.4.3 cover topics from remote sensing literature: speckle filtering, mutual information (an alternative to correlation), and manual selection of control points. Target recognition, bathymetry, InSAS, repeat-pass imaging, and change detection are important sonar applications and are summarised in Sections 4.4.4–4.4.7. Section 4.4.8 highlights important differences between sonar images and optical images. Section 4.4.9 reviews existing feature-based SAS works and Section 4.4.10 gives a sample of other SAS works. Finally, several developments in feature matching and registration of speckled imagery are described in Section 4.4.10.

4.4.1 Speckle filters

Since feature detectors are mainly designed for optical images, some SAR works propose the use of a speckle filter (or despeckling filter) to reduce the impact of speckle noise prior to feature detection and matching. Some of the simple filters include the mean (average) and median filters, while there are several adaptive statistical filters as well as some wavelet filters. A few of the notable filters include the Refined Lee filter [Lee 1981], the Frost filter [Frost et al. 1982], the anisotropic diffusion filter [Perona and Malik 1990][Yu and Acton 2002], and the wavelet filter [Jin et al. 2012]. Although speckle reduction methods are well considered in both radar and ultrasound [Wu et al. 2013], their specific use in registration applications is nonstandard and the benefits are fairly subjective. Speckle noise can be reduced more effectively using averaging techniques known as multi-look algorithms [Maître 2013][Huang and van Genderen 2014], though at the cost of image resolution.

Schwind et al. [2010] recommended using a denoising filter [Shen and Castan 1992] before using SIFT, and [Midtgaard et al. 2011] used anisotropic diffusion prior to feature detection. The SIFT algorithm [Lowe 2004] assumes the original image has a Gaussian blur sufficient to prevent aliasing and performs smoothing of the image before building the scale-space representation. Prior smoothing, whose amount is parameterised as σ_0, serves to increase the repeatability of features. In computer vision, simple blurring filters such as Gaussian blur are often applied to remove noise prior to image processing. For registration, these filters may increase the precision rate (i.e., percentage of correct matches in the chosen set) at the cost of decreasing the recall rate (percentage of correct

matches out of all possible correct matches) and reducing the localisation accuracy of keypoints [Shen and Castan 1992][Schwind et al. 2010].

4.4.2 Mutual information

Whereas correlation captures linear dependence, mutual information [Cover and Thomas 1991] is a measure of dependency that captures non-linear dependence. Mutual information can be considered a generalised measure of similarity that indicates how much information (or reduction of certainty) one variable provides about another. Mutual information is measured in bits and yields unbounded non-negative values (from 0 to ∞). Zero mutual information means that two random variables are independent, whereas a high mutual information indicates a high dependence. Mutual information can also be normalised to form normalised mutual information [Knops et al. 2006] (with values ranging from 0 to 1) or other normalised variations [Bouma 2009].

Mutual information can be used to register multi-spectral, multisensor, or multimodal images in spite of their different radiometric properties, such as the registration of SAR and optical images [Inglada 2002][Shu and Tan 2007]. Aside from robustness across radiometric differences, mutual information can produce sharper peaks than correlation and therefore provide an improvement in localisation accuracy, but it is more expensive to compute [Inglada 2002].

4.4.3 Manual control points

In some remote sensing applications, control points (also known as tie points) are manually selected by a human instead of finding point correspondences using automatic methods [Bentoutou et al. 2005][Dellinger et al. 2015]. Suitable control points may include landmarks or easily identifiable structures; these control points are identified and localised manually. This approach becomes infeasible when a large number of images need to be registered. Another consideration that affects appropriate selection of image registration algorithms is whether each image registration can be modelled with local transformations (where image warping may be suitable) or a global transformation model is required (to create a consistent global registration) [Brown 1992]. A semi-automatic control point algorithm was proposed by Kennedy and Cohen [2003].

4.4.4 Target/object recognition

Mine hunting applications are a prominent area of interest in sonar. Ideally, an automatic detection and classification (ADAC) system is able to detect and classify mines without requiring the participation of any human operators for interpreting the images [Leier 2014]. This includes the sub-problem of automatic target recognition (ATR), a term that generally refers to the detection of mines. In some applications, detected

mines should also be referenced against past information [Gendron et al. 2007] or be mapped in a more global positional context [Fallon et al. 2013]. Real-time functionality may also be desired [Gendron et al. 2007]. A common scheme for ADAC is to perform image segmentation, followed by feature extraction then classification. Segmentation aims to separate and identify regions of the image as object, shadow, and background [Fandos 2012]. Statistical and geometrical features are used to characterise the segmented regions; one such set of features was proposed by Fandos et al. [2013]. (Note that features here refer to classification rather than image features.) Many choices of classifiers are available.

4.4.5 Bathymetry and InSAS

Bathymetry is the 2.5D reconstruction of the seabed topography. A traditional technique for surveying the seafloor is to use a ship-mounted echo sounder, which measures the depth directly beneath the ship. To survey an area, a series of runs along parallel lines must be made over the scene. Multibeam systems improve on the efficiency of mapping by utilizing a set of echo sounders pointing radially at various angles orthogonal to the direction of travel, providing greatly increased coverage [de Moustier 1988][Calder and Mayer 2003]. An alternative method is to estimate depth values in a scene by using the sonar data from two (or more) closely spaced receivers; this is referred to as bathymetric side-scan sonar [Denbigh 1989]. Both of these systems fall under the category of swath bathymetry, where the area of seafloor that is insonified is a fan-shaped swath below the sonar.

Interferometry refers to the use of the coherent echo signals at separate hydrophones to measure the angle of a returning wave from a sonar target. Thus, a bathymetric side-scan sonar (such as a system with two vertically-displaced hydrophones) applies the following principle of bathymetric reconstruction: the relative time delay of echoes measured at the vertically-displaced hydrophones is estimated as accurately as possible; the angle of arrival of a signal can be computed from the time delay, and subsequently a height estimate is derived from the angle given the system geometry. This type of interferometry is called across-track interferometry, where the baseline or displacement between the sensors is perpendicular to the look direction of the sensor and the along-track axis. The accuracy of bathymetric estimation is dependent on the baseline, where too short a baseline yields poor estimates of height and too long a baseline leads to spatial baseline decorrelation and larger phase errors [Li and Goldstein 1990]. With InSAS, synthetic aperture processing simplifies an array of vertically-displaced hydrophones into a single element.

4.4.6 Repeat-pass sonar

Interferometry can also be performed over repeat-pass images [Bellec et al. 2005], although it is less accurate for bathymetry since the relative displacement of the transducers between runs is not known as accurately. Repeat-pass interferometry is a special case of along-track interferometry [Rosen et al. 2000], which maps the spatio-temporal coherence of the scene and also allows for change detection, ground moving target indication, and deformation estimation. As with SAS, there are several practical challenges to InSAS and repeat-pass interferometry such as navigation, motion compensation, handling occlusions, ambiguities due to phase wrapping, and multipath. Repeat-pass interferometry is more challenging than InSAS due to knowledge of the baseline being less accurate and the inevitable influence of temporal decorrelation between runs.

4.4.7 Change detection

Change detection is the identification of temporal differences between multiple images of the same scene. This can be useful for applications such as harbour surveys. There are two main categories of change detection: object-based change detection and image-based change detection. In object-based change detection, also known as contact-based change detection and feature-based change detection, targets are first identified in the new image, then the set of targets is compared with historical records to determine new and missing targets as well as positioning information. Object-based change detection has some overlap with target recognition, and often requires an operator to manually identify objects. Object-based change detection requires the general signature of targets to be known in advance. It is a form of supervised detection in that it relies on data that has been labelled or classified, possibly also requiring labelled ground-truth training data, whereas unsupervised detection requires no prior information or context other than the images themselves [Bruzzone and Prieto 2000]. With image-based change detection, which is typically unsupervised, the output is a difference image that indicates the degree of change for each pixel in the scene. Image-based change detection is more sensitive to changes in illumination, viewing geometry, and imaging artefacts (which should not be detected as changes), and generally requires actively navigated platforms [Myers et al. 2014]. Figure 4.1 shows a difference image computed from the pixel-wise subtraction of two coregistered despeckled log-intensity images. The middle image is the repeat-pass image after four objects were deliberately placed in the scene, and the changes due to these objects are clearly visible in the difference image.

Automatic change detection is a topical application in the SAS field and is particularly of interest for military surveillance of coastal seafloors. It is also relevant for offshore oil and gas exploration, pipeline surveying, and oceanography [Dillon 2013]. Change detection is performed via image differencing and thresholding, which may be based on measures of similarity and/or entropy. Image-based change detection

can be divided further divided into coherent methods and incoherent methods. Incoherent change detection uses only the magnitude of the complex speckle images (or even greyscale images) and can be effective over longer intervals between repeat passes, whereas coherent change detection requires shorter intervals but can detect subtle changes in the scene (such as a patch disturbed by a mine burial) even when the mean backscattered energy remains the same [Myers et al. 2017]. Per-pixel changes can be determined on the basis of correlation/coherence or more sophisticated measures such as canonical correlation analysis [G-Michael et al. 2016] and temporally invariant saliency [Matthews and Sternlicht 2011]. Image registration is a precursor to change detection, with coherent change detection being especially reliant on an accurate registration in order to produce useful results. The most common rule of thumb states that a registration to within a tenth of a resolution cell is required for interferometric processing, estimation of changes in topography, and coherent change detection [Just and Bamler 1994][Scheiber and Moreira 2000][Preiss and Stacy 2006]. Although this requirement originates from SAR, it has also been applied to SAS [Sæbø et al. 2011][Dillon and Myers 2014b]. Misalignment between two images causes spurious false alarms in image-based change detection [Myers et al. 2009]. Incoherent change detection can be performed under looser conditions for image coregistration and navigation estimation accuracy.

4.4.8 Differences between SAS images and optical images

Synthetic aperture imagery differs from optical imagery in several respects. Pixel values of speckled images are complex-valued rather than colour or greyscale. Image resolution depends on the bandwidth of the transmitted wave. Transducers propagate waves that illuminate the scene, also causing shadows. In SAS these shadows may be particularly blurred due to the projected waves originating from different angles as the sonar runs along a track [Sabel et al. 2005][Bellettini and Pinto 2009]. The transmitted energy is usually the only source of illumination whereas in optical scenes there are often multiple sources of lighting that may change in luminance and direction [Marburg et al. 2012]. SAS imagery is also subject to potentially severe distortion due to the sonar system swaying and drifting from an ideal track. The extent of image warping can be reduced if the track navigation data is known accurately. In practice, the data from inertial measurement units (IMUs) and GPS tracking can only be used as initial estimates; path correction relies on the recorded sonar data instead. Speckled images tend to be texturally bland, whereas optical images tend to have a variety of distinctive textures. Image noise in speckled imagery is largely multiplicative in nature, as opposed to additive noise in optical images. Synthetic aperture images are often interpreted (by both humans and computers) on a log-magnitude scale showing an intensity range up to 60 dB.

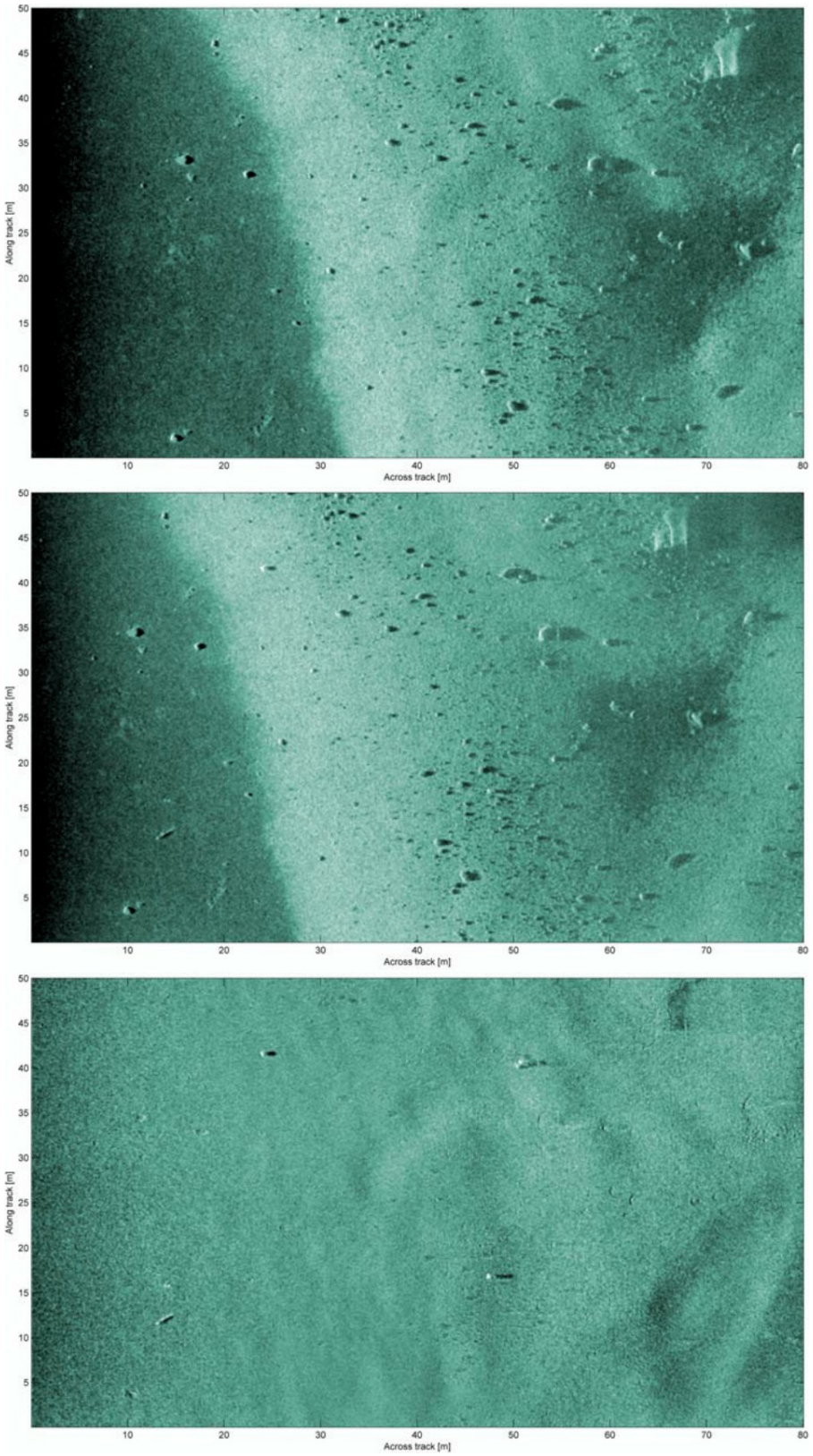


Figure 4.1: SAS images of a scene before and after deployment of four target objects (top and middle images respectively). The bottom image is the difference image, which shows the visible changes due to the objects. The dynamic range in these images is 45 dB.

There are several important differences between how SAR and SAS imagery is processed, as well as the maturity of the technical literature. SIFT and its alternatives are mentioned in SAS research only recently [Wang and Hayes 2014] and in passing [G-Michael et al. 2014][Midtgaard et al. 2011], whereas its application is more deeply considered [Schwind et al. 2010][Wang et al. 2015][Suri et al. 2010][Dellinger et al. 2015][Wang et al. 2012][Fan et al. 2013] in the better established SAR community.

4.4.9 Feature-based SAS work

At the time of writing, there are only a few independent papers that explore the feasibility of feature-based image registration for SAS. However, feature-based registration has been gaining popularity in SAR and can in theory be applied to SAS in a similar manner. A key difference between SAS and SAR is the difficulty of accurate positioning for SAS. The strict requirements on navigation accuracy of a sonar platform usually cannot be met through IMUs or other external positioning systems alone and requires data-driven techniques [Dillon and Myers 2014b]. Furthermore, motion errors in water are comparable to the signal wavelength and can have a highly destructive effect on image formation algorithms [Caporale and Petillot 2017]. The primary motivation behind feature-based registration for SAR and SAS to employ a faster initial registration to reduce the search space of a subsequent area-based method that is slower but more accurate.

G-Michael et al. [2014][2016] performed initial coarse registration (for alignment to within one pixel, ideally) by estimating a 2D image shift (no rotation or scaling) from SIFT feature matches. They used a threshold of two thirds for Lowe's distance ratio. Their estimate was a weighted average based on descriptor distances, and outliers were rejected by repeatedly culling data points further than three standard deviations from the sample mean. (However, random sampling algorithms such as RANSAC [Fischler and Bolles 1981] are the more common approach to outlier rejection in registration problems, including for SAR applications [Dellinger et al. 2015]). Another SAS work using feature-based registration was by Midtgaard et al. [2011]; they used anisotropic diffusion to reduce speckle as a preprocessing step. Correspondences found using SURF were used to estimate an affine transform between the images from two sonar runs. Unfortunately, neither of these works presented results indicating the performance of the feature algorithms. [Midtgaard 2013] also used SURF to estimate an affine transform prior to performing pixel-wise subtraction for image-based change detection. 3000 feature matches were found between two images taken four days apart, whereas pairs of images taken about one year part produced 11 and 16 matches each.

4.4.10 Other SAS work

Repeat-pass applications are still a relatively new undertaking in SAS research. Bellec et al. [2005] described an early attempt at repeat-pass interferometry, obtaining a maximum correlation near 0.76 and using two mine echoes as reference points to align the images. Many SAS papers involving AUV navigation and mine detection have no references to computer vision or even image processing. Similarly, many publications of practical applications give little detail on the inner workings of their systems. Fallon et al. [2013] described an application where an AUV with a forward looking sonar used a simultaneous localisation and mapping (SLAM) algorithm [Durrant-Whyte and Bailey 2006]. Gendron et al. [2007] described a real-time mine detection system with SLAM-like navigation. Lyons et al. [2010] modelled the effect of seafloor ripples on SAS speckle statistics.

Lyons and Brown [2013] performed experiments using a rail-mounted tower system where repeat-pass tracks achieved a navigation accuracy surpassing that of normal practical scenarios. Their findings included an analysis of temporal decorrelation for a scene that changed relatively quickly due to bioturbation caused by feeding fish. Their example is unique in that the coherence was particularly high due to the accuracy of their registration; this accuracy cannot be achieved in practice and is typically only possible in simulations. Dillon and Myers [2014a] performed a brief feasibility study of vertical baseline estimation for repeat-pass interferometric SAS, suggesting that a small baseline gives ambiguous estimates while large baselines yield decorrelated views. A universal issue pointed out was the inability to measure a ground truth vertical separation, as the state of the art instrumentation is insufficient. [Myers et al. 2009] attempted automatic change detection and suggested the use of local co-registration for regions of suspected change in order to reduce spurious false alarms due to global misalignments.

4.4.11 Other developments in feature matching and registration of speckle images

Some SAR papers report surprisingly low number of feature correspondence inliers ($< 10\,\%$) whereas other works mention over $90\,\%$ inliers [Wang et al. 2012], perhaps indicating the highly situational nature of feature matching performance. [Ren et al. 2011] used feature matching with restrictions based on scale and rotation discrepancies between potential feature matches. [Hasan et al. 2010] considered a weak form of guided matching for multi-spectral image registration. In the case of a high inlier ratio in non-real-time applications, there is little incentive to improve on readily available SIFT implementations, especially with overall insignificant objective benefit demonstrated by its variants. For example, no works have considered and demonstrated an improvement over SIFT's keypoint accuracy, and few consider the effect on accuracy of estimated

image registration parameters. Several works lowered the false alarm rate of SIFT at the cost of keypoints, which inevitably decreases the computation time. Such a trade-off is arguably trivial. An extreme case of misusing the false alarm rate metric was evident in one work where the majority of keypoints were redundant (though not exactly identical). A contrario variants of RANSAC (AC-RANSAC) are suitable when the inlier ratio is low (even 10 %) [Moisan and Stival 2004] and have the general advantage of having a single parameter [Dellinger et al. 2012]. The a contrario model assumes either a uniform [Moisan and Stival 2004] or Gaussian [Sur 2010] distributed localisation error and uses the expected number of false alarms (NFA) at a 5 % null hypothesis significance to determine outliers.

In feature matching, dual matching is defined as requiring a correspondence pair of keypoints to pass Lowe's distance ratio test both ways [Wang et al. 2012]. [Huo et al. 2012] implemented a multilevel SIFT matching process in a coarse-to-fine matching approach for efficiency and to deal with the problem of ambiguity in very high resolution (VHR) images. [Fan et al. 2013] combined several minor SIFT tweaks appearing in the SAR discussion and also concatenated SIFT descriptors of sizes 16×16, 24×24, and 32×32 into a single descriptor in attempt to maximise robustness, yielding an improvement in the inlier ratio of matches.

Registration of images from different sensing technologies (e.g. SAR, infrared, optical, CT, MRI) can be performed using a technique called pixel migration [Keller and Averbuch 2006][Yao and Goh 2006], where no explicit similarity measure is required between the multisensor images. Yao and Goh [2006] noted that local optimisation of the implicit similarity function does not give the most accurate registration, adapting the problem using a genetic algorithm that searches for appropriate feasible and infeasible regions to register.

Chapter 5

Feature-based SAS baseline estimation

This chapter describes a proof of concept of feature-based registration for SAS imagery. A feature matching pipeline using SIFT is proposed, with a novel approach to sonar image registration via sonar track registration. The method is demonstrated on two simulated SAS images with slightly differing aspects (and no temporal changes), where the ground truth registration is exactly known due to its relation to the specified simulation inputs. It is shown that SIFT performs well on these simulated speckled images and leads to a registration result within the 0.1 pixel alignment accuracy recommended for coherent change detection. For the problem of track registration, a simple track model is used to consider the feasibility of SIFT in ideal conditions without further refinement using area-based methods. An estimated track registration is used to construct an image registration, i.e., a mapping of points between the pair of images. One motivation for track registration rather than direct image registration is the potential for the estimated track registration to lead to improved motion compensation, image formation, and thus better change detection results (for example).

Section 5.1 presents the imaging model for an ideal sonar track. Section 5.2 derives equations for estimating track parameters from a set of image correspondences. Section 5.3 details a least squares estimation method that provides improved estimation accuracy. Section 5.4 describes the use of RANSAC for outlier rejection and proposes a deterministic procedure that approximates the behavior of RANSAC. An overview of the test data and the image preprocessing involved is given in Section 5.5. Section 5.6 presents the results and performance of feature matching on the dataset. Section 5.7 demonstrates the performance of RANSAC estimation. The results of the proposed registration pipeline after least squares estimation is given in Section 5.8. Finally, Section 5.9 discusses the various results and ideas for further improvement and development.

5.1 Imaging model for an ideal sonar track

This section presents a registration model for two sonar tracks. Consider a seafloor scene imaged in two separate SAS runs. The sonar runs are assumed to be ideal—along straight and level tracks, where the relative pose of the sonar on the track at any given sampling time has no pitch, roll, yaw, sway, surge, or heave. A point scatterer with world coordinates ${}^{w}\mathbf{x}$ will appear in the local coordinate systems of each run as $\mathbf{x}$ and $\mathbf{x}'$, where

$$\mathbf{x} = \mathbf{R}\ {}^{w}\mathbf{x} + \mathbf{t}\ , \tag{5.1}$$

$$\mathbf{x}' = \mathbf{R}'\,{}^{w}\mathbf{x} + \mathbf{t}'. \tag{5.2}$$

Here, the rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$ describe the overall pose of the first run, and $\mathbf{R}'$ and $\mathbf{t}'$ describe the pose of the second run. The local coordinates of a scatterer can be parameterised for each run as

$$\mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \sqrt{r^2 - z^2} \\ y \\ z \end{pmatrix}\ , \tag{5.3}$$

$$\mathbf{x}' = \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \sqrt{r'^2 - z'^2} \\ y' \\ z' \end{pmatrix}; \tag{5.4}$$

where r and r' are the slant ranges, x and x' are the across-track positions, y and y' are the along-track positions, and z and z' are the vertical displacements of the scatterer with respect to each of the two sonar tracks. After synthetic aperture reconstruction at track altitude, an estimate is obtained of the slant range and along-track position but not the depth; the scatterer appears in the two images at the coordinates $\mathbf{r}$ and $\mathbf{r}'$, where

$$\mathbf{r} = (r, y)^{\mathrm{T}}\ , \tag{5.5}$$

$$\mathbf{r}' = \left(r', y'\right)^{\mathrm{T}}. \tag{5.6}$$

Without loss of generality, assume that the second run is the reference so that $\mathbf{R}' = \mathbf{I}$ and $\mathbf{t}' = \mathbf{0}$, and thus

$$\mathbf{x}' = {}^{w}\mathbf{x} = \mathbf{R}^{-1}(\mathbf{x} - \mathbf{t}), \tag{5.7}$$

$$\mathbf{x} = \mathbf{R}\,\mathbf{x}' + \mathbf{t}. \tag{5.8}$$

If there is no pitch or roll between the tracks, the rotation matrix has the form

$$\mathbf{R} = \begin{bmatrix} \cos\alpha & \sin\alpha & 0 \\ -\sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{5.9}$$

where α is the angle of rotation between the first and second track orientations. The translation vector $\mathbf{t} = (t_x, t_y, t_z)^{\mathrm{T}}$ represents the displacement going from the first track to the second track relative to the world coordinate system. Using (5.7), (5.8), and (5.9), the mapping of points between the two track coordinate systems is given by:

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x'\cos\alpha + y'\sin\alpha + t_x \\ -x'\sin\alpha + y'\cos\alpha + t_y \\ z' + t_z \end{pmatrix}; \tag{5.10}$$

where the inverse mapping is

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} (x - t_x)\cos\alpha - (y - t_y)\sin\alpha \\ (x - t_x)\sin\alpha + (y - t_y)\cos\alpha \\ z - t_z \end{pmatrix}. \tag{5.11}$$

Note that the depth of a scatterer in a scene generally cannot be measured to a sufficient accuracy for registration (without bathymetry); only the image coordinates are available. In addition, if the sonar tracks and seafloor are not horizontal, the mapping of image points can be mathematically ambiguous. Thus, the seafloor and the sonar tracks are assumed to be level.

Equations (5.10) and (5.11) are the projected coordinates for a given track registration and are used in the next section to measure the misregistration for an estimated track registration.

5.2 Track registration from image correspondences

Assume two level tracks over the same level scene (see Figure 5.1), where the sonar altitude above the seabed, H', of the second sonar track is known or estimated. The track registration problem is to estimate the relative pose between the tracks, given a set of N image correspondences $\mathbf{r}_k \leftrightarrow \mathbf{r}'_k$, $k \in 1..N$, where $\mathbf{r}_k$ and $\mathbf{r}'_k$ are the matched image coordinates (5.5) and (5.6) that are related by (5.8). It is assumed that there are a sufficient number of good correspondences, otherwise registration using image features may not be feasible. This assumption depends on various factors such as the image quality, baseline, decorrelation, and scene content. Note that if no altitude (or scene

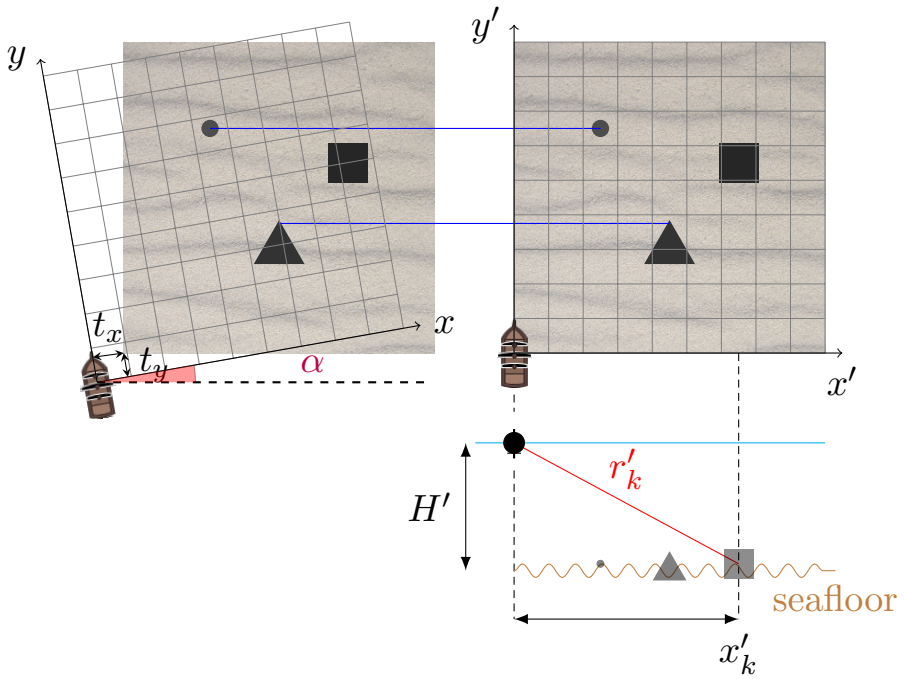


Figure 5.1: The first and second tracks are depicted on the left and right respectively, along with the scene. The first track has horizontal displacements t_x and t_y relative to the second coordinate system. The orientation of the first track also has a counter-clockwise rotation, α, relative to the second track, as well as a vertical separation, t_z, that is not shown here. The bottom right shows the across-track geometry for a point scatterer with slant range r'_k and a known depth H'. The two blue lines are examples of feature correspondences, where the same points in the scene are found in both images. Reprinted, with permission, from [Wang and Hayes 2017b]. © 2017 IEEE.
Use of "Sand texture" by Rowland Rose, 2009, via Flickr (CC BY 2.0).

depth) is known, there are an infinite number of possible depths and baselines that are consistent with the image data such that an arbitrary solution for track registration still leads to a useful image registration. If both track altitudes (H and H') are known, the system of equations presented in this section becomes overdetermined. This scenario is considered in Appendix A, but ultimately, estimates of both track altitudes are unlikely to be measured accurately enough to be of practical use.

The set of image correspondences are inevitably noisy: correspondences will be imperfectly localised due to finite processing on finite resolution images, and some incorrect correspondences may be found using SIFT. If all the correspondence outliers are eliminated, track parameter estimation can be defined and solved as an optimisation problem.

For a given track registration parameterised by $\mathbf{t}$ and α, the projected image co-

ordinates based on (5.10) and (5.11) respectively are:

$$\tilde{\mathbf{r}}_k = \begin{pmatrix} \sqrt{(-H' + t_z)^2 + \left(x'_k \cos\alpha + y'_k \sin\alpha + t_x\right)^2} \\ -x'_k \sin\alpha + y'_k \cos\alpha + t_y \end{pmatrix} \quad , \tag{5.12}$$

$$\tilde{\mathbf{r}}'_k = \begin{pmatrix} \sqrt{(-H')^2 + \left((x_k - t_x)\cos\alpha - (y_k - t_y)\sin\alpha\right)^2} \\ (x_k - t_x)\sin\alpha + (y_k - t_y)\cos\alpha \end{pmatrix}. \tag{5.13}$$

The errors $\Delta\mathbf{r}_k$ and $\Delta\mathbf{r}'_k$ are the differences between the correspondence locations and their predicted locations for a given registration estimate:

$$\Delta\mathbf{r}_k = \mathbf{r}_k - \tilde{\mathbf{r}}_k = \begin{pmatrix} r_k - \tilde{r}_k \\ y_k - \tilde{y}_k \end{pmatrix}, \tag{5.14}$$

$$\Delta\mathbf{r}'_k = \mathbf{r}'_k - \tilde{\mathbf{r}}'_k = \begin{pmatrix} r'_k - \tilde{r}'_k \\ y'_k - \tilde{y}'_k \end{pmatrix}. \tag{5.15}$$

The symmetric transfer error (measured in units of distance squared) for a correspondence $\mathbf{r}_k \leftrightarrow \mathbf{r}'_k$ is

$$\overline{E}_k = \frac{\|\Delta\mathbf{r}_k\|_2{}^2 + \|\Delta\mathbf{r}'_k\|_2{}^2}{2}. \tag{5.16}$$

Then, the average symmetric transfer error [Hartley and Zisserman 2003] for a set of N correspondences is

$$\overline{E} = \frac{1}{N}\sum_{k=1}^{N} \overline{E}_k. \tag{5.17}$$

Thus, given a correspondence set $\mathbf{r}_k \leftrightarrow \mathbf{r}'_k$, $k \in 1..N$ and assuming a level seafloor at depth H below the first sonar track, the optimisation problem is defined as:

$$\hat{\alpha}, \hat{\mathbf{t}} = \arg\min_{\alpha,\mathbf{t}} \left\{\overline{E}\right\}. \tag{5.18}$$

A naïve approach to this problem is to perform a brute-force search for an approximate global minima of (5.17), followed by a local optimisation of this average symmetric transfer error. However, this is slow due to the large search space and the shape of the error function. An alternative to brute force is to estimate the global minima by direct calculation of the parameters using correspondences. With reference to Figure 5.1, the derivation below demonstrates that the four track parameters are completely determined by a pair of non-identical correspondences, $\mathbf{r}_1 \leftrightarrow \mathbf{r}'_1$ and $\mathbf{r}_2 \leftrightarrow \mathbf{r}'_2$. From (5.11), the difference between two along-track positions relative to the second track is

$$y'_1 - y'_2 = (x_1 - x_2)\sin\alpha + (y_1 - y_2)\cos\alpha, \tag{5.19}$$

which is solved for α by:

$$\alpha = 2\tan^{-1}\left(\frac{x_1 - x_2 \pm \sqrt{(x_1-x_2)^2+(y_1-y_2)^2-(y_1'-y_2')^2}}{y_1'-y_2'+y_1-y_2}\right). \tag{5.20}$$

This yields two solutions for α, only one of which is consistent with the subsequent solutions. Next, the world coordinate offset,

$$m = -t_x\cos\alpha + t_y\sin\alpha, \tag{5.21}$$

is solved as follows. From the first row of (5.4):

$$r_1'^2 - r_2'^2 = x_1'^2 - x_2'^2.$$

Substituting the first row of (5.11) yields

$$\begin{aligned} r_1'^2 - r_2'^2 &= (c_1+m)^2 - (c_2+m)^2, \\ &= (c_1^2 - c_2^2) + 2(c_1-c_2)m, \end{aligned}$$

where

$$c_1 = x_1\cos\alpha - y_1\sin\alpha, \tag{5.22}$$

$$c_2 = x_2\cos\alpha - y_2\sin\alpha, \tag{5.23}$$

which leads to the solution

$$m = \frac{r_1'^2 - r_2'^2 - c_1^2 + c_2^2}{2(c_1-c_2)}. \tag{5.24}$$

Then x_1' is solved from (5.22) and (5.24):

$$x_1' = c_1 + m; \tag{5.25}$$

and finally the track registration parameters are solved by:

$$t_x = x_1 - x_1'\cos\alpha - y_1'\sin\alpha, \tag{5.26}$$

$$t_y = y_1 + x_1'\sin\alpha - y_1'\cos\alpha, \tag{5.27}$$

$$t_z = -H + \sqrt{r_1'^2 - x_1'^2}. \tag{5.28}$$

Thus, given a non-degenerate pair of inlier correspondences, an estimate of the track parameters can be calculated directly.

5.3 Estimation using least squares

Although a single pair of inlier correspondences can be sufficient to form an estimate of the repeat-pass sonar baseline, more reliable and accurate estimates can be formed by fitting to a larger set of correspondences, reducing the overall effect of noise in the localisation of the correspondences. With homography estimation in computer vision, a standard method is to perform local optimisation on the average symmetric transfer error [Hartley and Zisserman 2003]. However, the sonar geometry differs from the epipolar geometry of cameras, and so the equivalent error function, (5.17), is somewhat unstable with noisy data due to the presence of many local minima. In general, initial estimates computed directly from pairs of correspondences are not reliable enough to be refined either. The problem of convergence on sub-optimal solutions can be addressed by using more expensive optimisation algorithms. For example, an initial global brute-force search can be used prior to local optimisation on the error function. An alternative approach proposed here is to use a least-squares formulation in order to estimate the track registration more reliably and efficiently. By definition, the least squares solution gives slightly different results to minimisation of (5.17). The least squares regression consisting of two equations is derived as follows.

Firstly, taking the second row of (5.11) and using 3rd order Taylor expansions to approximate $\sin\alpha$ and $\cos\alpha$ results in:

$$\begin{aligned} y' &\approx (x - t_x)\left(\alpha - \frac{\alpha^3}{6}\right) + (y - t_y)\left(1 - \frac{\alpha^2}{2}\right), \\ &\approx (x - t_x)\alpha - \frac{\alpha^3}{6}(x - t_x) + y - t_y - \frac{\alpha^2}{2}(y - t_y). \end{aligned}$$

This equation is cast as a local linear approximation to yield the first linear regression equation:

$$y' - y + \frac{\hat{\alpha}^2}{2}(y - \hat{t}_y) + \frac{\hat{\alpha}^3}{6}(x - \hat{t}_x) = (x - \hat{t}_x)\alpha - t_y, \tag{5.29}$$

where α and t_y are solved using vectors of the variables, given initial estimates of t_x, t_y, and α ($\hat{t}_x$, $\hat{t}_y$, and $\hat{\alpha}$ respectively). After a few iterations where the estimates are updated, the solution converges.

Secondly, the following equation is linearised in order to estimate the remaining parameters, t_x and t_z:

$$r' = \sqrt{\hat{r}'^2}.$$

Substituting the 1st order Taylor polynomial for the square root about $\hat{r}'^2$ to obtain a local linear approximation gives:

$$r' \approx \hat{r}' + \frac{1}{2\hat{r}'}[r'^2 - \hat{r}'^2], \tag{5.30}$$

where $\hat{r}'$ is calculated using the estimates $\hat{t}_x$ and $\hat{t}_z$ according to the first row of (5.11):

$$\hat{r}' = \sqrt{(z - \hat{t}_z)^2 + \big((x - \hat{t}_x)\cos\hat{\alpha} - (y - \hat{t}_y)\sin\hat{\alpha}\big)^2}. \tag{5.31}$$

After expanding the expressions for r' (from (5.11)) and $\hat{r}'^2$, taking the difference, and replacing t_x and t_z with their estimates $\hat{t}_x$ and $\hat{t}_z$ as required for simplification, the result is linear in t_x and t_z:

$$\begin{aligned} r'^2 - \hat{r}'^2 \approx &- \left[\hat{t}_z^2 - 2z\hat{t}_z + (\hat{t}_x^2 - 2x\hat{t}_x)\cos^2\hat{\alpha} + \hat{t}_x(y - \hat{t}_y)\sin\hat{\alpha}\cos\hat{\alpha}\right] \\ &+ (\hat{t}_z - 2z)t_z + \left[(\hat{t}_x - 2x)\cos^2\hat{\alpha} + (y - \hat{t}_y)\sin\hat{\alpha}\cos\hat{\alpha}\right] t_x. \end{aligned}$$

Substituting this result into the earlier square root approximation, (5.30), yields:

$$\begin{aligned} &r' - \hat{r}' + \frac{1}{2\hat{r}'}\left[\hat{t}_z^2 - 2z\hat{t}_z + (\hat{t}_x^2 - 2x\hat{t}_x)\cos^2\hat{\alpha} + \hat{t}_x(y - \hat{t}_y)\sin\hat{\alpha}\cos\hat{\alpha}\right] \\ &\quad\approx \frac{1}{2\hat{r}'}\left[(\hat{t}_x - 2x)\cos^2\hat{\alpha} + (y - \hat{t}_y)\sin\hat{\alpha}\cos\hat{\alpha}\right] t_x + \frac{1}{2\hat{r}'}(\hat{t}_z - 2z)t_z. \end{aligned} \tag{5.32}$$

As a linear regression equation, $\hat{r}'$ is the main source of noise in the predictor expressions and should be eliminated by multiplying both sides by $\hat{r}'$ to give:

$$\begin{aligned} &\hat{r}'(r' - \hat{r}') + \frac{1}{2}\left[\hat{t}_z^2 - 2z\hat{t}_z + (\hat{t}_x^2 - 2x\hat{t}_x)\cos^2\hat{\alpha} + \hat{t}_x(y - \hat{t}_y)\sin\hat{\alpha}\cos\hat{\alpha}\right] \\ &\quad\approx \frac{1}{2}\left[(\hat{t}_x - 2x)\cos^2\hat{\alpha} + (y - \hat{t}_y)\sin\hat{\alpha}\cos\hat{\alpha}\right] t_x + \frac{1}{2}(\hat{t}_z - 2z)t_z. \end{aligned} \tag{5.33}$$

Ordinary least squares is used to solve the linear regressions. The two regressions in (5.29) and (5.33) are solved iteratively and alternately so that updated estimates of the parameter subsets interact with one another. (A single linear regression can be formulated but this is tedious and unnecessary.) In the first linear equation, (5.29), the response variable is roughly equal to the across-track correspondence offset, since the rotation (α) is kept small in multi-pass sonar applications. Contrary to the least-squares assumption of noise-free predictors, x (appearing in the predictor for α) will have a non-zero error in relation to feature detection. x is also part of a predictor in the second linear equation, (5.33). However, experiments using weighted least squares (based on descriptor distances) and total least squares (where errors in the independent variables are taken into account) yielded concordant solutions.

Although ordinary least squares is robust to noisy data, it is not robust to outliers. Hence, rejection of outlier correspondences must be performed prior to least squares estimation in order to provide robustness to bad estimates and degenerate cases. Overall, this least-squares approach provides more accurate and more stable estimation than local optimisation and is considerably faster to compute.

5.4 Outlier rejection using RANSAC

In the ideal case, a track registration can be estimated from a minimal set of two correspondences. In practice, however, a given set of image correspondences will have noisy feature locations and contain outlier matches. Furthermore, inlier pairs of correspondences can form degenerate pairs in the given calculations. These issues are typical in pose estimation problems, where RANSAC is a common choice of outlier rejection algorithm used with feature matching [Torr and Zisserman 2000]. The track registration model described in Section 5.2 was to used to estimate the track parameters by using (5.16)—the symmetric transfer error (measured in squared pixels)—as the RANSAC reprojection error. In practice, the one-way errors (5.14) and (5.15) are always similar in value, and so the square root of the RANSAC threshold (referred to as the "tolerance" here) behaves as the maximum allowable Euclidean distance error for the predicted mapping of a supposed inlier correspondence. Note that the SIFT detector can detect multiple keypoints at the same location with different orientations [Lowe 2004], so redundant point correspondences were removed before performing robust estimation using RANSAC.

Although RANSAC estimation is robust, it is known to be unstable in terms of estimation accuracy [Chum et al. 2003]. When the inlier ratio is high, RANSAC tends to offer little benefit since its emphasis is on finding an inlier set rather than minimising the effect of noise on estimation. In comparison, the least-squares method is designed to minimise estimation error in the presence of noise. Using least squares estimation for fitting tentative RANSAC inlier sets does not solve this issue because RANSAC ranks solutions by the cardinality of the inlier set [Fischler and Bolles 1981], not any measure of error or quality of fit.

As demonstrated in the results in Section 5.7, inlier rejection and noise minimisation should be treated as separate concerns when robustness is required but accuracy is of primary importance. Based on this idea, a modified outlier rejection scheme is proposed here.

5.4.1 A deterministic approximation to RANSAC

Although RANSAC is useful for coarse outlier rejection, for finer thresholds it is prone to finding inlier sets that yield biased but unstable estimates. In order to provide similar but more stable and predictable results, the following deterministic algorithm is proposed for approximating RANSAC's behavior for smaller error thresholds.

An iteratively trimmed inlier set is obtained for a given reprojection tolerance as follows:

1. The current registration estimate is assumed to be sufficiently accurate in terms of being used to cull one outlier from the current inlier set, using the same error

measurement as in RANSAC.

2. A new registration estimate is formed (to ensure minimal bias).

Schwind et al. [2010] refer to a similar ad-hoc method as "filtered matching", and [G-Michael and Tucker 2010] also uses a similar idea. Prior to this iterative rejection scheme, RANSAC is still used for (nondeterministic) coarse outlier rejection in order to ensure that the initial data is not degenerate and the first estimate is reasonable. However, when the threshold is set in order to discard less accurate but technically inlier correspondences, RANSAC should be used with a pre-determined coarse threshold (such as 1.0 squared pixels) before applying the proposed iterative algorithm to cull points to the desired smaller threshold. Although multiple outliers (as opposed to just one) can be culled in each iteration, the number removed each iteration should be kept small so that each removal does not have an unstable impact on the updated estimate.

5.5 Test data

The simulated high frequency SAS data was generated by the NSWC PCD (Panama City, Florida) using PC-SWAT [Sammelmann 2001]. The scene consists of three highly reflective point targets on a sandy seabed with ripples. The scene is imaged twice along two tracks that differ only by an altitude separation of 0.2 m, which is the known ground truth. Therefore, the simulated registration parameters are:

$$\begin{pmatrix} t_x & t_y & t_z & \alpha \end{pmatrix} = \begin{pmatrix} 0 & 0 & -0.2 & 0 \end{pmatrix}. \tag{5.34}$$

The two images are reconstructed at their respective track altitudes using the wavenumber reconstruction algorithm [Cafforio et al. 1991][Milman 1993], resulting in images of size 1800×1658. The known pose between these tracks is used to compute the ground truth mapping between the two images, which in turn is used to objectively evaluate the localisation accuracy of feature correspondences and overall registration accuracy.

5.5.1 Sonar image preprocessing

To perform feature matching, the complex sonar images reconstructed at track altitude must be converted to greyscale images. This was done by taking the log-magnitude images limited to a chosen dynamic range (with the maximum value set to 0 dB and small values clamped to a minimum intensity), followed by scaling these values to the integer range 0 to 255. Figure 5.2 shows a converted greyscale image with a dynamic range of 40 dB. The intensity profile in this image varies significantly over the slant-range direction due to the vertical beam patterns of the transducers and the decrease in signal amplitude with range. Although Figure 5.2 is trimmed to a minimum 5 m slant

range (since the region below is less than the sonar track altitude), this information was not presumed.

Preliminary testing revealed that SIFT yielded more inlier feature matches when operating on intensity-normalised images, where all the rows of each image (the pixels in a row share the same slant range) are scaled to have approximately the same average intensity. This normalisation is performed on the complex images, not the greyscale images. There are multiple ways to achieve a satisfactory normalised image; the simplest method of dividing the rows by their mean intensity may not give the best result because a single bright point can account for as much as 10 % of the total energy of its row. The sharp variations using this technique contrast with the true beam pattern and attenuation with distance, which are both smooth functions. For this reason, an estimate of the median intensity was used instead of the mean intensity. Model-based approaches could also have been used to normalise the images.

Local features are more effective when the illumination or intensity in a scene is more consistent, rather than having areas of low contrast. Making the images more uniform resulted in more detected features (twice as many features at a dynamic range of 40 dB) with better spread and a marginally narrower error distribution in terms of localisation accuracy; all of these factors theoretically decrease statistical estimation bias. Figure 5.3 shows one of the converted greyscale images after intensity compensation. Note that this image also has a dynamic range of 40 dB, but the intensity scale is not the same as in Figure 5.2; a higher dynamic range is required for the intensities to be similar for the brightest regions of the unnormalised image. In general, an appropriate choice of dynamic range depends on the system parameters as well as the scene data.

The chosen dynamic range of converted images is relevant to SIFT performance as it affects the level of contrast of features and therefore the number of detected features. When the dynamic range is low, there are fewer bright pixels, and when the dynamic range is high, there are fewer dark pixels. At a certain sweet spot, SIFT finds the most correspondences. SIFT is partially invariant to changes in the level of illumination, and preliminary experiments showed that the dynamic range had a trivial impact on the localisation accuracy of correspondences. A 30 dB dynamic range for the normalised images was chosen for the experiments due to providing enough correspondences while requiring less computation than higher dynamic range images. Using the 40 dB normalised images produced four times as many correspondences with approximately the same mean error and standard deviation in localisation error (compared to ground truth), however, a raw increase in the number of correspondences did not necessarily lead to a more accurate registration. Although a factor of two increase in accuracy was found to be possible using a dynamic range between 40–50 dB compared to 30 dB, there was still a significant element of randomness. The processing time of SIFT matching is roughly quadratic in the number of detected features due to brute-force matching.

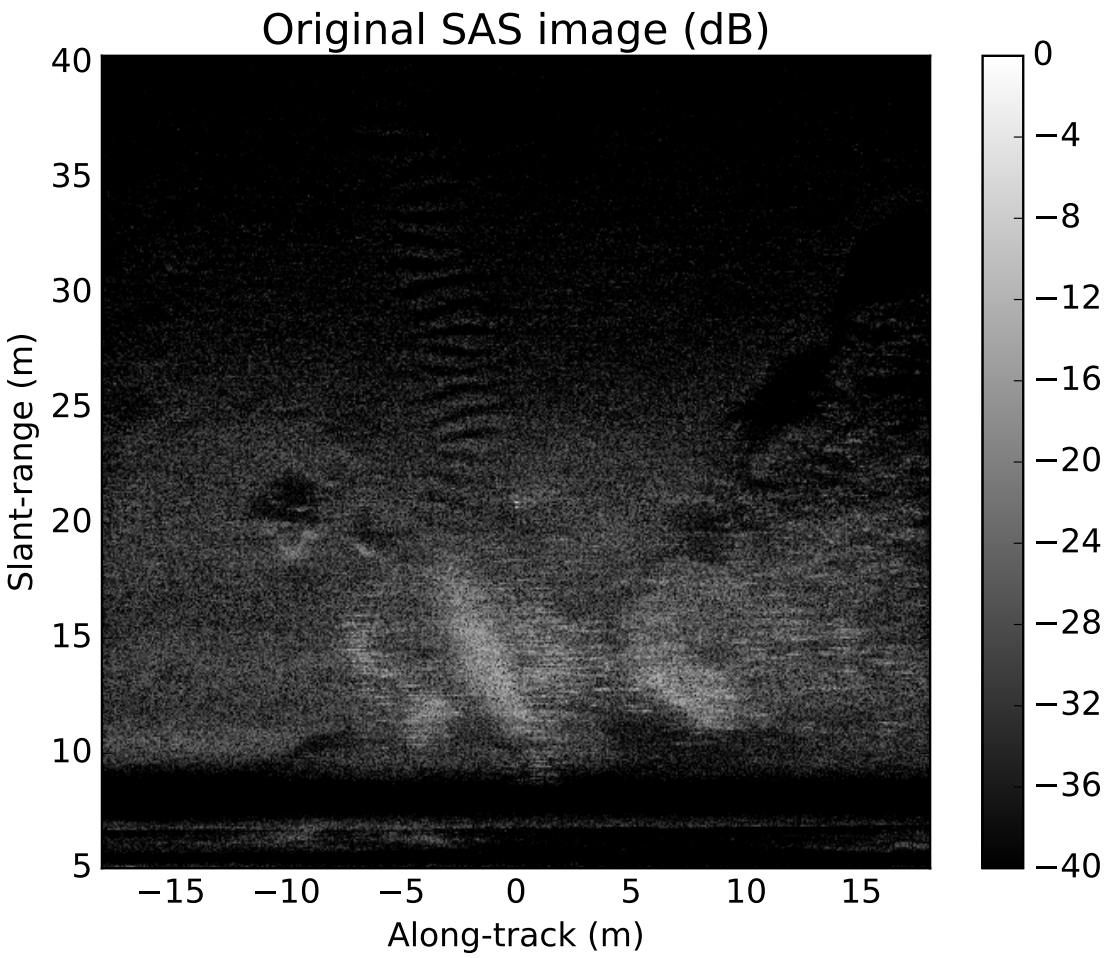


Figure 5.2: The original SAS image reconstructed at track altitude from the first sonar run. The along-track axis is horizontal with slant range increasing upwards. The dynamic range shown is 40 dB.

5.6 Feature matching performance

Feature detection, description, and matching were performed on the two SAS images using the SIFT algorithm implemented in the OpenCV library [Bradski 2000]. The SIFT parameters used were as recommended in Lowe's original paper [Lowe 2004] except for the contrast threshold, which was set to 0.04. The results of SIFT matching are listed in Table 5.1. Redundant correspondences due to SIFT keypoints with different orientations at the same location were removed. The repeatability of the keypoints is about a quarter, and there are very few outliers in the matches. Although 0.5 was used for the distance ratio test, the precision rate remained steady for all ratios below 0.9. In preliminary tests, the SIFT detector and descriptor achieved the best distribution of features in terms of localisation accuracy compared to other combinations of detectors and descriptors, including SURF. Other publically available implementations of SIFT were also tested, with inferior results. The recently popular KAZE detector/descriptor

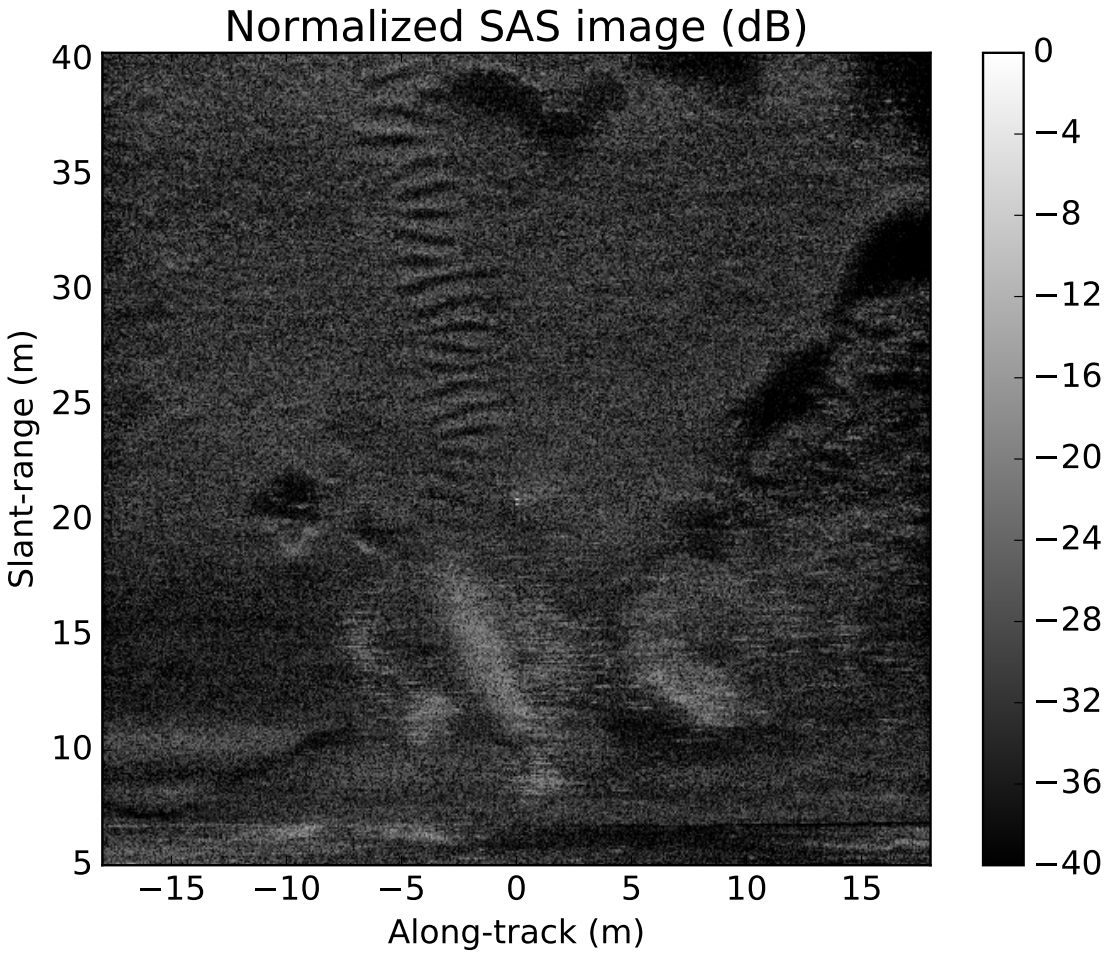


Figure 5.3: The intensity-normalised image for the first sonar run, with a dynamic range of 40 dB. The bottom region where the slant range is less than 5 m is unused. © 2017 IEEE.

and its derivative A-KAZE [Alcantarilla et al. 2013] appeared impractical for use on large images at the time of testing, being prohibitively slow, requiring large amounts of memory, and yielding unimpressive results.

The ground truth mapping of points between the two images was determined from the known sonar tracks. Using the ground truth, the localisation accuracy can be easily observed based on the along-track pixel offsets of the correspondences, since the two simulated sonar runs have no along-track displacement. The slant-range error offsets were also computed. The along-track and slant-range offsets were combined to form the Euclidean distance error (in pixels) of the correspondences relative to the known ground truth. Histograms of the along-track offsets are shown in Figure 5.4, the slant-range offsets in Figure 5.5, and the distance error in Figure 5.6. The mean and standard deviations of these populations are listed in Table 5.2. A correspondence was defined as an outlier if either of its point locations differed from their predicted ground truth projections by greater than 1 pixel L_∞-norm distance. (In side exper-

iments where the correspondences were ranked according to descriptor distances and distance ratios, there was no observed relationship between the closeness of a match and its localisation accuracy. Thus, it did not seem possible to use this information to improve estimation using weighted estimation. There was also no statistically significant correlation between the magnitude of a localisation error and the slant range of the matched points.)

Table 5.1: SIFT performance.

SIFT keypoints in first image	21,093
SIFT keypoints in second image	16,710
Unique-location matches	4901
Inlier unique-location matches	4893
Precision rate (proportion of inlier matches)	99.8 %

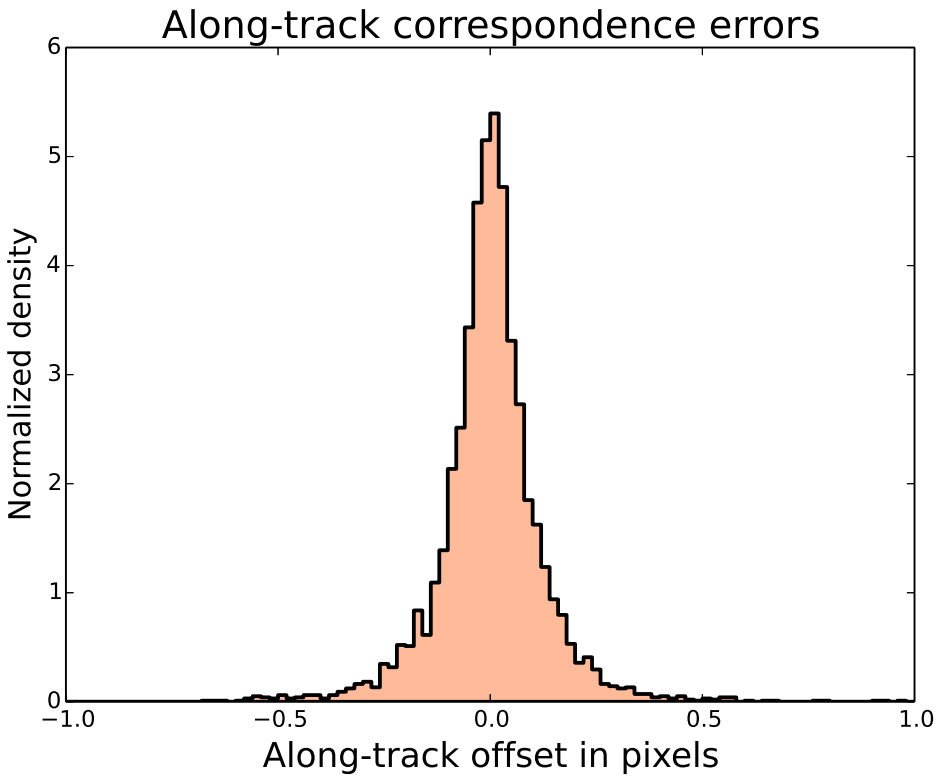


Figure 5.4: Normalised histogram of the along-track pixel offsets between corresponding feature locations, compared to the ground truth. Outliers are not shown. In the absence of speckle noise, an ideal feature detector would always yield offsets close to zero.

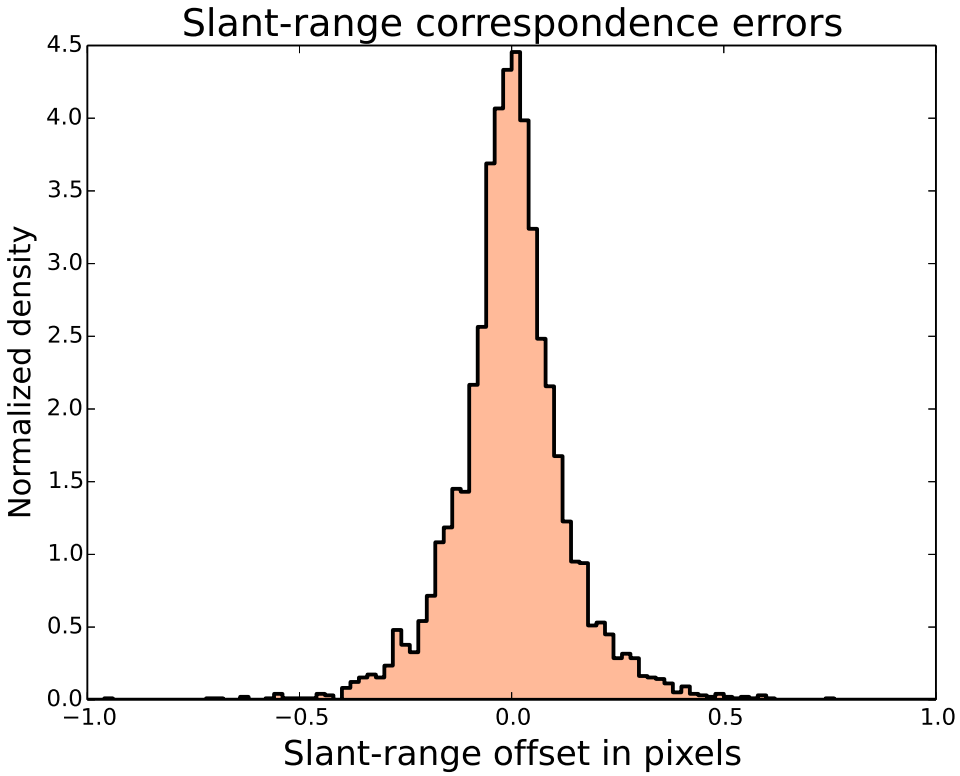


Figure 5.5: Normalised histogram of the slant-range pixel offsets between corresponding feature locations, compared to the ground truth. Outliers are not shown.

Table 5.2: Correspondence error statistics.

Statistic	mean	stddev
Along-track error (pixels)	0.00051	0.125
Slant-range error (pixels)	−0.0027	0.131
Distance error (pixels)	0.140	0.115

5.7 RANSAC estimation performance

RANSAC was used to compute track registration estimates at various error thresholds up to 1.0 pixels. For the minimal set of two correspondences, the initial parameter estimates were obtained using the method of direct calculation (see Section 5.2). In the case of more than two correspondences, model fitting was performed using the least squares regression from Section 5.3 rather than optimisation of the symmetric transfer error. This choice was made for efficiency reasons. For each RANSAC solution, a least squares fit was applied to the largest inlier set to obtain the final estimate. The quality of a track registration estimate was measured by the maximum registration error it produced over the scene when its predictions were compared to the ground truth. Due to RANSAC's non-deterministic nature and its random sampling scheme,

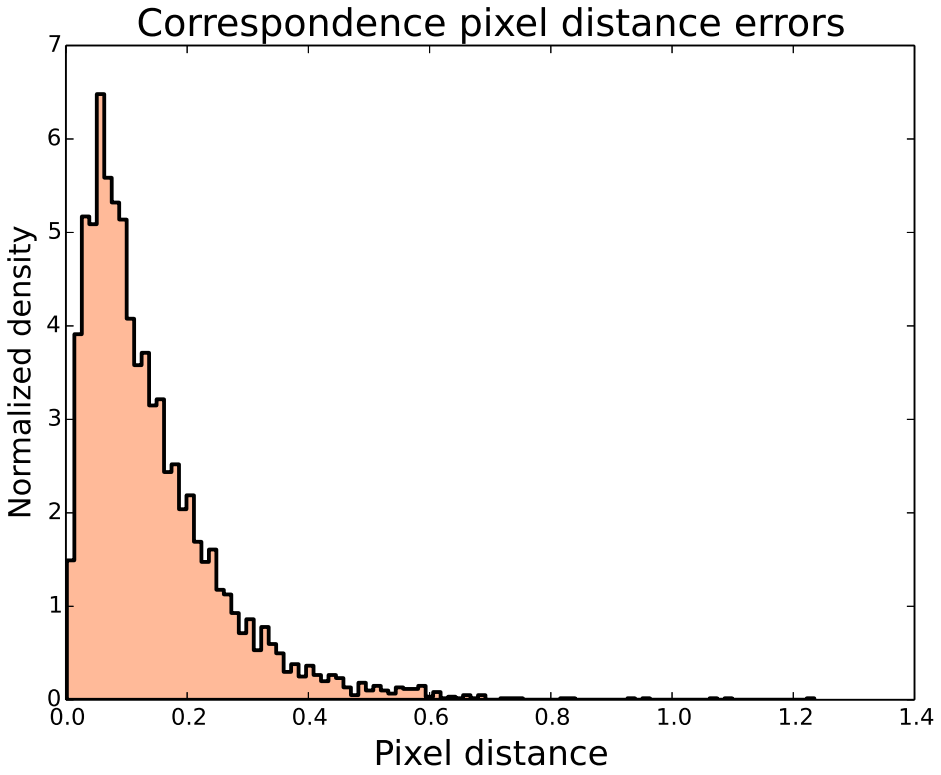


Figure 5.6: Normalised histogram of the Euclidean pixel distance error between corresponding feature locations compared to the ground truth. Outliers are not shown.

its performance in terms of registration accuracy cannot be (usefully) characterised statistically. For example, it is extremely unlikely but possible for RANSAC to fail to find a solution as long as there is at least one outlier correspondence in the data set. Therefore, it is not meaningful to consider a maximum error bound. Similarly, the best possible estimate that RANSAC can find is an arbitrary value determined by a brute-force combination of subsets of the data.

However, for demonstrative purposes, populations of 20 RANSAC solutions (found from 10,000 iterations each) were obtained at each of the thresholds: 1.0^2, 0.1^2, and 0.05^2 pixels squared. The minimum and maximum registration errors and sizes of the inlier sets are displayed in Table 5.3. For a distance tolerance of 1.0 pixels, there is a high probability of finding the maximum possible inlier set (of 4895) in all of the 20 runs. (Distance tolerance is equivalent to the square root of the RANSAC threshold.) With the two smaller tolerances (0.1 and 0.05), the misregistration errors varied significantly. These examples are consistent with the well-known behavior of RANSAC where too large a threshold leads to poor estimates [Torr and Zisserman 2000] and too small a threshold leads to unstable estimates.

The most reliable representation of RANSAC's output is perhaps a single sample corresponding to the largest inlier set that can be found given a sufficient (i.e., infinite)

number of iterations. This solution can also be computed using brute force, but it is prohibitively expensive to do so. Instead, an iteration count of 10,000 was used under the assumption that this was sufficient to approximate RANSAC's "ideal" behavior. Sample performances of RANSAC using this approach are plotted in Figure 5.7, which shows the effect of varying the RANSAC threshold. The large variation in quality of sample estimates is apparent as the pixel distance tolerance decreases. Minima in this plot have no significance as they cannot be repeatedly obtained. There does not appear to be an ideal choice of threshold that would make a noticeable impact on registration accuracy, especially in a practical scenario where the ground truth would not be known.

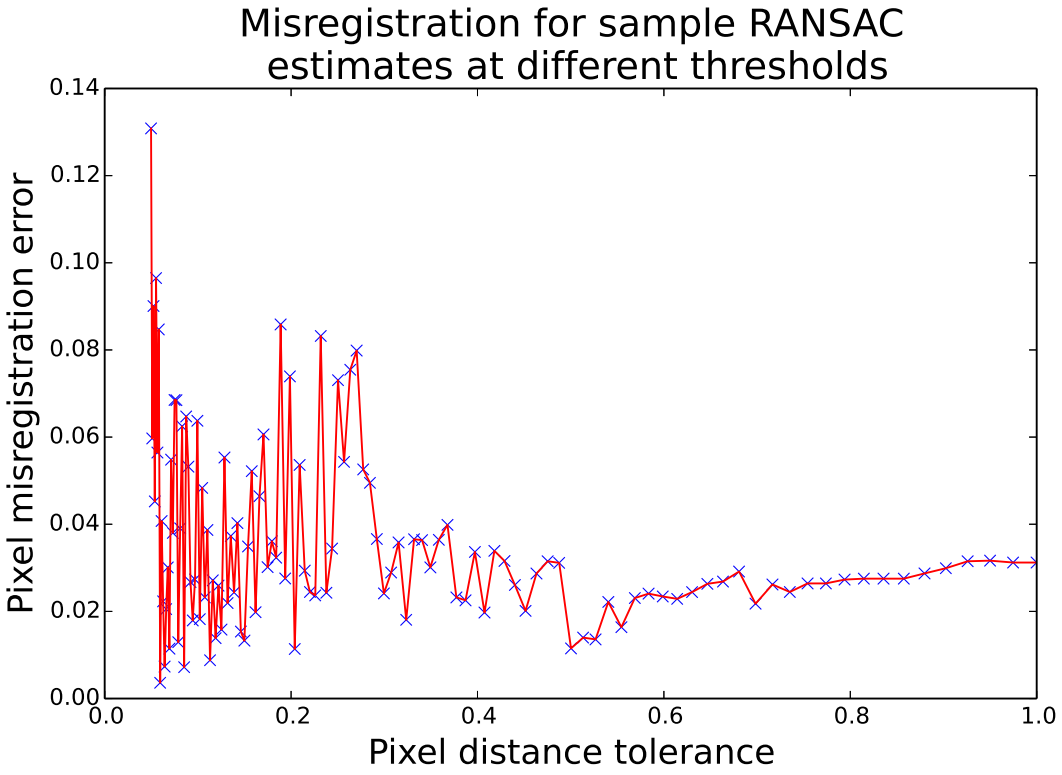


Figure 5.7: Misregistration corresponding to sample RANSAC solutions (found from 10,000 iterations each) with the RANSAC threshold varying up to 1.0 pixels squared. For better visibility, the horizontal axis uses the pixel distance tolerance, which is the square root of the RANSAC threshold.

Table 5.3: Example populations of RANSAC estimates.

Threshold	Number of inliers		Misregistration (pixels)	
(pixels^2)	min	max	min	max
1.0^2	4895	4895	0.031	0.031
0.1^2	2265	2315	0.008	0.094
0.05^2	924	950	0.043	0.191

Next, Inlier sets of size 2 to 4890 were built using the ground truth. Only correspondences whose Euclidean distance error was less than a specified tolerance were considered inliers. Figure 5.8 shows the relationship between the chosen tolerance and the misregistration of the estimate computed from ground-truth inliers; these results are deterministic for a given data set. Estimation was performed using the proposed least-squares formulation, which tends to give slightly better results. As with the trend in Figure 5.7, the misregistration also becomes more chaotic with fewer inliers. Although the global minima is ten times less than the 0.028 pixel error obtained using an a priori choice of 1.0 pixels tolerance, it seems there is no continuous range of tolerances that would be reliably superior for another similar dataset. For this reason, we recommend a RANSAC threshold of 1.0 pixels squared for outlier rejection and suggest that there is no real downside to this choice when the ground truth is unknown.

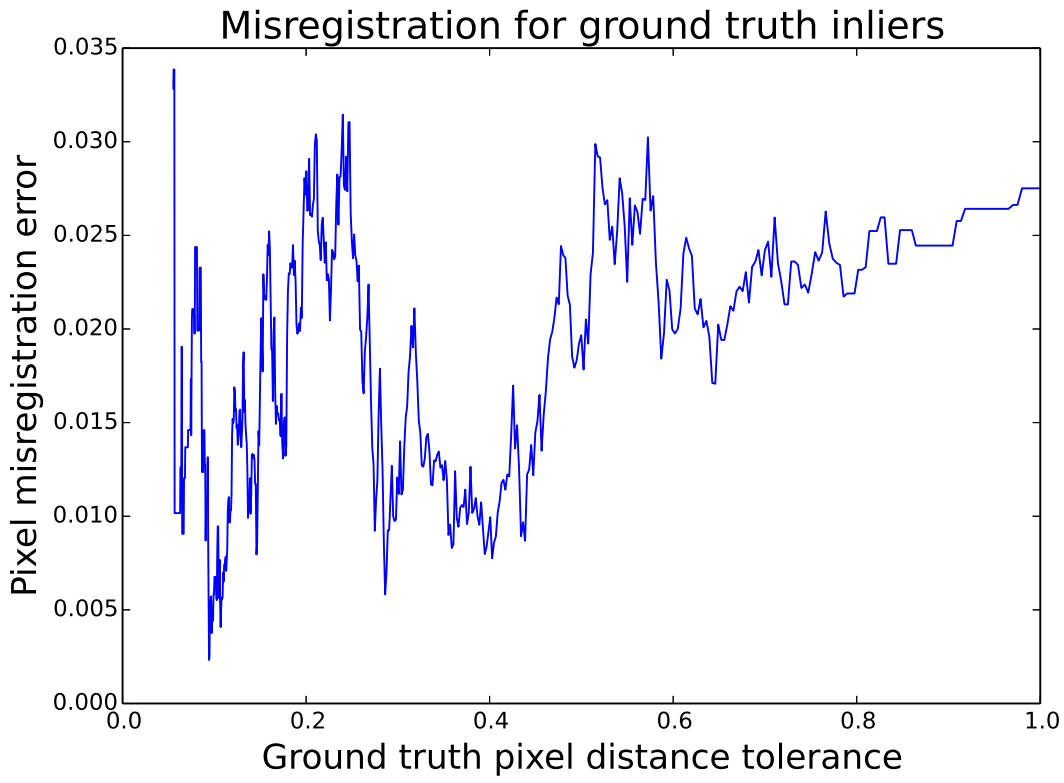


Figure 5.8: Misregistration corresponding to estimates from ground-truth-based inlier sets, where inlier correspondences have a Euclidean distance error less than a specified tolerance. The lack of clear trend implies that automatic selection of a near-optimal error threshold (as one might hope for when using RANSAC) cannot be performed reliably without the ground truth.

5.8 Least-squares registration performance

Firstly, robust outlier rejection was performed using RANSAC with a threshold of 1.0 pixels squared. 4895 inliers were found from 4901 correspondences. Next, using the same error threshold of 1.0, the worst inliers were iteratively discarded using the method proposed in Section 5.4.1. Six more correspondences were rejected in this process. (At the same threshold, the proposed method often rejects additional outliers. This is because RANSAC finds the solution that maximises the number of inliers, which differs from the concept of minimising the error fit of the solution.) Least squares regression resulted in the estimated parameter values in Table 5.4 (c.f. (5.34)).

Table 5.4: Estimated parameters using ordinary least squares regression.

$\hat{t}_x$ (m)	$\hat{t}_y$ (m)	$\hat{t}_z$ (m)	$\hat{\alpha}$ (°)
1.13×10^{-4}	-5.49×10^{-5}	-0.1993	1.26×10^{-4}

The estimated registration parameters were used to predict the mapping of scene points in the two images, according to (5.10) and (5.11). The scene was sampled in a uniform grid and the predicted scene-point mappings compared to the ground truth mapping in order to calculate the alignment error (in pixels) over the scene. This misregistration is depicted in Figure 5.9. In this case, the misregistration is worst close the sonar track and better at a greater slant range. The quality of the registration is deemed to be based on the maximum misregistration error (0.028) despite most of the scene being registered to within 0.01 pixels. (In the case of using 40 dB images instead of 30 dB images, 19,326 feature matches were found and the estimation procedure led to a least squares solution with a maximum registration error of 0.012 pixels. Although it is likely that the 40 dB images would give better results overall, the long computation time for computing plots such as Figure 5.7 and Figure 5.8 to demonstrate representativeness of accuracy was somewhat impractical.

5.9 Discussion

The results shown in Figures 5.4, 5.5, and 5.6 confirm that SIFT maintains sub-pixel accuracy on the given speckled SAS imagery, with 47 % of correspondences being localised to within 0.1 pixels of the ground truth. In the absence of speckle noise, an ideal feature detector would be expected to consistently yield errors close to zero. The distributions of the along-track and slant-range offsets are both roughly symmetrical with approximately zero mean and a sharp peak. They do not seem to follow any known distributions and attempts to model them were unsuccessful.

The results are overall promising. The least squares estimation led to a registration error up to 0.028 pixels over the scene, which is within the 0.1 pixel alignment accu-

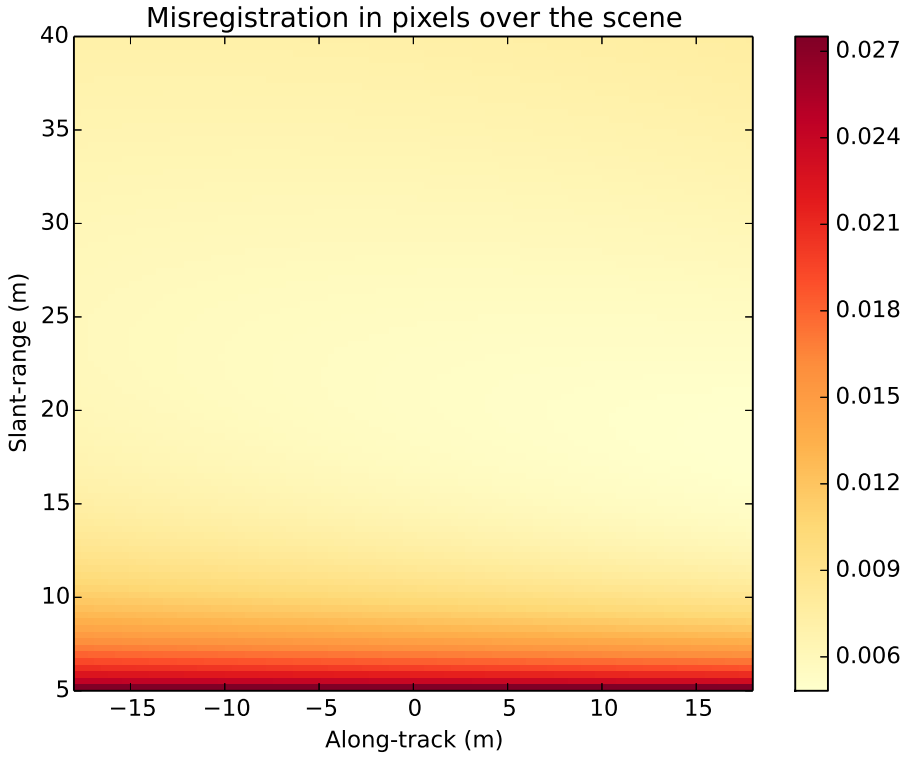


Figure 5.9: The alignment error of the estimated registration compared to the ground truth.

racy required to compute a high quality interferogram [Scheiber and Moreira 2000]. Although RANSAC is designed for robustness and not specifically for reducing the effect of noise, the variation in the quality of RANSAC estimates was slightly unexpected. Using a threshold of 0.1^2 pixels squared rather than 1.0^2 pixels squared conserved nearly half of the inliers, especially the inliers that supposedly had significantly less correspondence offset error. However, any search performed without using the ground truth knowledge is prone to bias, but even using the ground truth to retain only the correspondences with smaller localisation errors did not reveal a predictable or exploitable trend. This seems to imply that including correspondences with error offsets larger than 0.1 pixels does not necessarily decrease estimation accuracy but provides stability.

The relationship between the correspondence errors and the maximum misregistration is difficult to evaluate. For example, suppose that the distribution of the correspondence errors was known and the error distribution of the parameter estimate could also be modelled after regression. Even so, it becomes impractical to continue applying an error model due to the complex nature of the misregistration function, which is the local maximum of a symmetric reprojection error, (5.16). For this reason, robust estimators such as M-estimators are unlikely to provide a tangible benefit. The

use of popular RANSAC variants such as MLESAC [Torr and Zisserman 2000] and LO-RANSAC [Chum et al. 2003] are also unlike to make a difference because with few inliers, the problem is primarily of estimation, not outlier rejection or computational efficiency. Appendix B presents a statistical argument on why it may not be possible to construct confidence intervals for a set of parameters estimated from feature matches.

A possible adaptation is to use 40 dB images instead of 30 dB images, which results in four times as many detected features with a statistically similar distribution of localisation error. Theoretically, estimation from more points results in increased stability and decreased bias/sensitivity to noise. However, the running time increases by about 16 times (according to the quadratic time complexity of brute-force matching) in obtaining a final result with half the misregistration error. It is arguable whether this trade-off is worthwhile, and further testing with more datasets is required to determine whether a reduction in misregistration can be attained in general. Additionally, it is desirable to further improve the registration using a correlation-based method, where adapting a global feature-based registration model to a more nuanced model with local distortions could diminish the utility of a factor-of-two improvement in the initial misregistration accuracy.

The height offset t_z appears to be the most sensitive parameter in terms of both the parameter estimation error and the corresponding effect on the track misregistration. One might consider the scenario where both scene depths are estimated (see Appendix A for the resulting problem). t_z could then be calculated as the difference between the track altitudes. However, the estimate for t_z would need to be on the order of 1 mm accuracy for a registration to within 0.1 pixels to be possible. Depth measurements of such accuracy are unattainable in practice [Hagen and Jalving 2008]. Although depth estimation has an important impact on the track misregistration, the sonar track model has a depth ambiguity such that there are (infinitely) many track registrations that correspond to the same image registration. In other words, any sensible value of H' can be used to compute a theoretically equivalent image registration. The accuracy of the estimate of depth is only significant in the context of using the estimated track registration to correct the navigation data and regenerate the sonar images.

The proportion of true inlier correspondences after matching with Lowe's ratio test is high ($> 95\,\%$) for any distance ratio below 0.9. Despite the presence of speckle, the simulated dataset is rather ideal. Real data can vary greatly in environmental aspects, technical specifications, and thus also image quality. Therefore, it is difficult to predict how successful the proposed pipeline would be for non-simulated high quality SAS imagery. Deviation from an ideal sonar track is inevitable in practice due to imperfect navigation. The sonar altitude may drift, or the assumption of a flat seafloor may not be appropriate. In such cases, the usefulness of an ideal track registration is limited, even with path-corrected images. Therefore, it is of interest in future work to investigate

the applicability of SIFT-based registration in the following conditions: with different sonar baselines, where the extent of speckle decorrelation varies; using simulated InSAS data and a non-flat seafloor; and using a piece-wise model of a sonar track with noisy motion estimates provided.

There are various proposed modifications to improve on SIFT's robustness to speckle [Suri et al. 2010][Dellinger et al. 2012][Dellinger et al. 2015][Wang et al. 2012]. Compared to SIFT, many of these achieve reduced computation times, higher repeatability, and lower false alarm rates at the cost of fewer matches. However, there have been limited ideas about improving on the localisation accuracy of SIFT keypoints and matches. For SAR and SAS, image registration using features is an option that can be used to significantly reduce the processing time of a subsequent finer registration based on correlation. Increasing the accuracy in the feature-based registration could potentially save an order of magnitude time overall. An improvement in detector statistics at the cost of less matches cannot be assumed to be an objective trade-off. Indeed, a trivial trade-off can be achieved with any feature detector by adjusting thresholds. A receiver operating curve (ROC) describing the percentage of correct matches retrieved in relation to the false alarm rate with varying parameters is also not an objective measure because one detector may have many correct but redundant features whereas a reasonable detector would find a set of features with better global coverage. Therefore, we propose registration accuracy to be of greater interest. Feature localisation cannot be readily improved by modifying the SIFT detector, but there are several other possibilities. Estimation of location uncertainty in SIFT keypoints could be used to weight the data for registration, thus favouring the more (probabilistically) accurate matches. Zeisl et al. [2009] demonstrate a minor improvement to this effect on optical images; perhaps the statistical properties of speckle could be factored in. G-Michael et al. [2014] used descriptor distances to weight the registration estimate, but the distance is meaningless for inlier matches. One idea is to refine the localisation of a correspondence using correlation over the local image regions [Suri et al. 2009]. Other options involve violating scale-space assumptions. Despeckling filters are designed to reduce speckle noise, but filters could be used to enhance scale-space structures for better localisation (at the cost introducing artefacts). For example, different features will be best localised at different levels of contrast or dynamic range. Perhaps the optimal dynamic range can be estimated, or the average keypoint location over different magnitude scales could be used. Lastly, the scale-invariance property is usually not needed for SAR/SAS applications; it is conceivable that a blob detector operating in a non-Gaussian scale space [Pauwels et al. 1995][Duits et al. 2004] could yield better localisation than the SIFT detector on speckled images.

Overall, a larger dataset with various track configurations is required to achieve a deeper analysis of the interaction between SIFT and speckle as well as a better sense of the estimation accuracy. Unfortunately, SAS imagery is not readily available (unlike

SAR imagery). SAS scene simulation is non-trivial and may typically require hours of processing time. However, simulation is essential in order to objectively measure feature-based registration performance due to the lack of precise ground truth with real data. The difficulty of accurate simulation is a plausible explanation for why localisation accuracy remains a neglected subject in the treatment of feature-based registration on synthetic aperture imagery in the literature.

Chapter 6

SIFT localisation accuracy on interpolated images

In signal processing, sinc interpolation is the ideal interpolant for bandlimited systems with adequately sampled data [Shannon 1949], where discrete-time interpolation [Cavicchi 1992] uses the Dirichlet kernel (also known as the periodic sinc function and the aliased sinc function). However, the question arises: what are the consequences if speckle images are not well sampled? This chapter addresses this topic specifically from the point of view of feature matching performance rather than the theoretical implications of sampling with a SAS system. Since feature detectors such as SIFT locate extrema within scale space (not just image space) to achieve scale invariance, the interpretation of values between data points cannot be concordant with interpolation methods such as sinc or linear interpolation.

Section 6.1 details the random generation of speckle images and pairs of images related by an interpolated shift. Various aspects of feature matching performance are measured and evaluated in Section 6.2. The results are explained and the implications are discussed in Section 6.3.

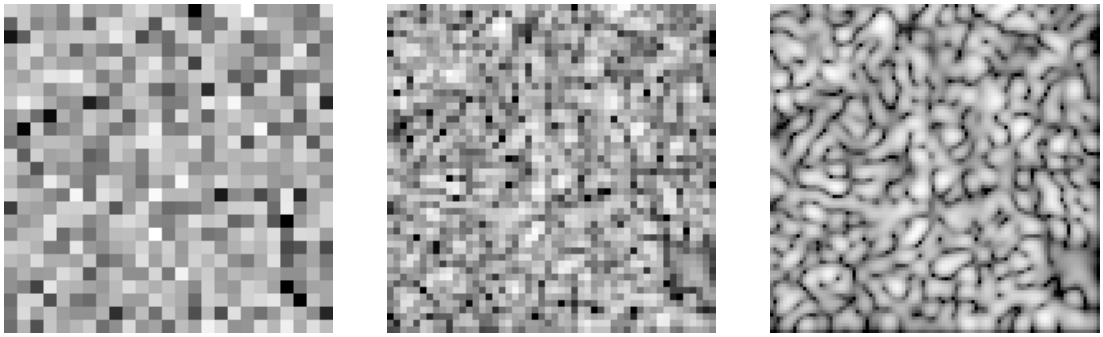
6.1 Generating speckle images with known ground truth

This set of experiments assumes fully developed speckle, which can be modelled as multiplicative complex Gaussian noise. For a perfectly bland scene that has a level surface with uniform reflectivity, the values of the resolution cells are independently and identically distributed and are drawn from a zero-mean unit-variance circular Gaussian distribution. This also assumes that: the speckle image is reconstructed at true depth; there is no aliasing or other source of noise; and the system is perfectly bandlimited so that all the samples in the images are uncorrelated. Complex speckle patterns are converted into a greyscale images by taking log-magnitude images with a limited dynamic range as described in Section 5.5.1. A 30 dB contrast was chosen.

When required, interpolation is performed on the speckle images prior to the greyscale conversion. To oversample images by an integer factor using sinc, images are zeropadded to twice the original dimensions (to avoid circular effects), then zeropadded

in the Fourier domain (increasing the doubled dimensions further by the interpolation factor), converted back to the time domain, and clipped to the correct dimensions in the time domain. Sinc interpolation can also be used to perform a fractional shift or delay of the signal (or more generally, arbitrary resampling) [Laakso et al. 1996]. A fractional shift of the image is performed by first zeropadding to twice the dimensions then multiplying by the linear phase $\exp(2\pi i f \Delta)$, where f is the digital frequency and Δ is the delay, followed by clipping to the correct dimensions in the time domain. For an image of size $M \times N$, a 2D shift requires multiplying by the linear phase $\exp(2\pi i f \Delta_x)$ across all M rows, as well as multiplying by $\exp(2\pi i f \Delta_y)$ across all N columns, where the frequency f is varying relative to the row or column being shifted. Note that to ensure a consistent interpretation, contributions to the Nyquist frequency bin should be split evenly between the 0.5 and -0.5 digital frequency bins. Using this resampling method, the ground truth relationship between two shifted images is known and can be precisely manipulated.

Figure 6.1a is an example of a random 25 px×25 px speckle pattern converted to a 30 dB greyscale image. Figures 6.1b and 6.1c show the same speckle pattern with oversampling factors of two and four respectively. With oversampling, the spatial structure of speckle and the speckle size [Dainty 1984][Fortune et al. 2004] become more apparent.



(a) No oversampling (b) 2× oversampling (c) 4× oversampling

Figure 6.1: Greyscale images of the same random speckle pattern under a) Nyquist sampling, b) 2× oversampling, and c) 4× oversampling. (These images show a dynamic range of 30 dB and have been scaled to the same image size for display purposes).

6.2 Feature matching performance on sinc-interpolated images using SIFT

Random speckle images of size 100×100 were generated as described in Section 6.1 according to desired sampling rates and fractional shifts. This small image size was chosen to reduce the computation time of brute-force matching (which is the main

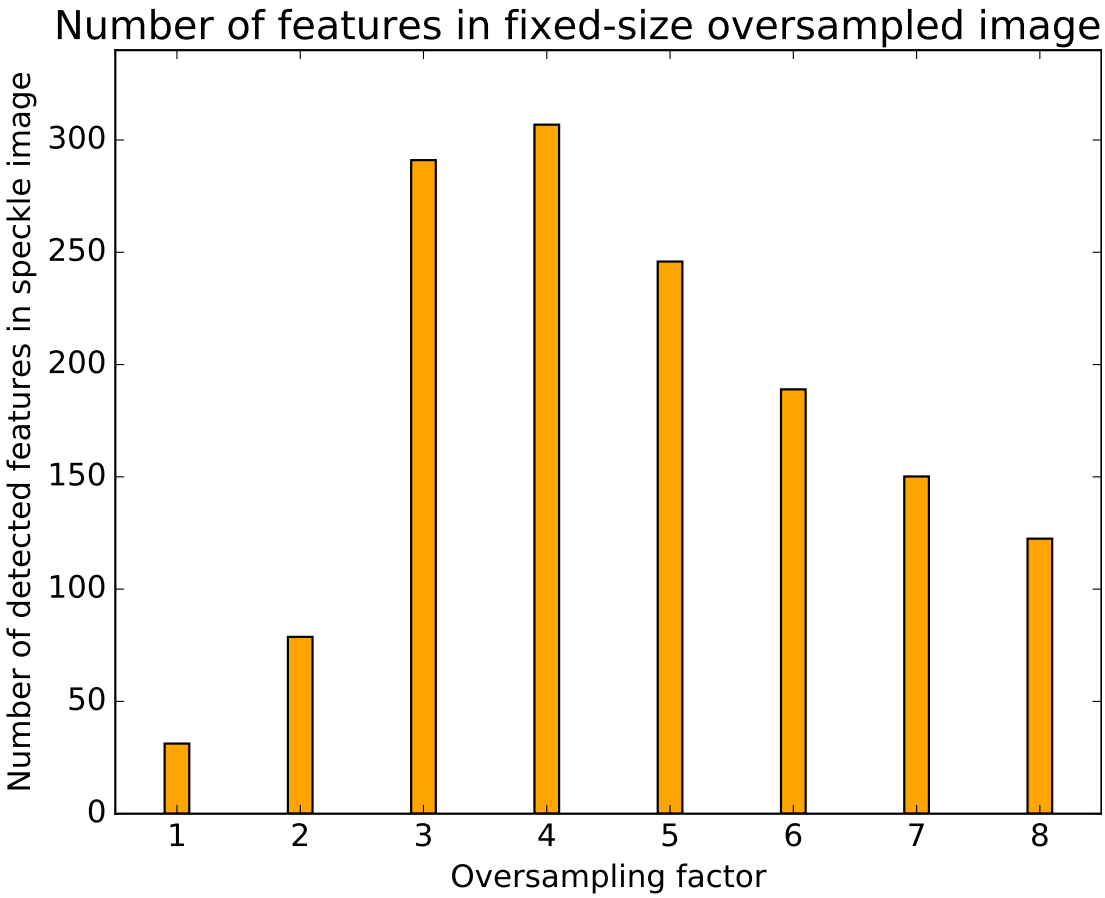


Figure 6.2: Oversampling by a factor of four yields the highest number of features relative to image dimension, about ten times the number of features for non-oversampled images (on average). These averages were computed from 1000 images.

bottleneck) and it has been confirmed that using larger images leads to insignificant differences in the resulting statistics. The OpenCV [Bradski 2000] implementation of SIFT was used to perform feature detection and description on pairs of greyscale images, followed by nearest-neighbour brute-force matching. Lowe's ratio test for removing weak matches was not performed. Instead, the known ground truth was used to remove outlier correspondences. A feature match was deemed an inlier if its two detected feature locations were localised to within one pixel Euclidean distance (after accounting for the controlled shift between the two images); otherwise, the match was rejected as an outlier. Although removal of outlier feature matches is not trivial in practice, this approach was used to measure the ideal performance of feature matching that a suitable outlier rejection method can approximate. (Furthermore, it is meaningless to consider the error statistics of outlier matches.)

In the first experiment, for each multiple of the Nyquist sampling factor ranging from one to eight, SIFT keypoint detection was performed on 1000 random images, resulting in the mean number of detected features per image shown in Figure 6.2.

Four-times oversampled images had the highest number of features on average—about ten times the number of features compared to Nyquist sampled speckle images. Since the oversampled images capture a smaller theoretical region of speckle given the fixed image dimensions, the number of detected features is also shown on a relative scale in terms of the physical area the image represents, as in Figure 6.3. The density of features increased significantly up to an oversampling factor of four, above which there is a diminishing gain with oversampling. Two-times, three-times, and four-times oversampling yielded factors of about 10, 84, and 157 times increases (respectively) in the feature density relative to no oversampling.

Next, the performance of SIFT was measured in relation to fractional shift amount along one axis using sinc interpolation. For each shift amount, 1000 random speckle images and their shifted image pairs were generated. Samples of the inlier feature matches (from the 1000 runs) were collated for each shift, with the shift amounts ranging from 0.0 to 1.0 pixels along the x-axis. Three main statistics were measured for each combined set: the ratio of inlier matches, the mean Euclidean localisation errors of the correspondences, and the standard deviation of these errors. Each of the performance curves were estimated for the cases of Nyquist sampling, two-times oversampling, and four-times oversampling, yielding the results shown in Figure 6.4 and Figure 6.5.

For each shift value, the mean signed error was also estimated from the 1000 images for the same chosen sampling factors (one, two, and four), with the result shown in Figure 6.6. Without oversampling, the bias curve has a sinusoidal shape, whereas with oversampling the bias values become too small to discern a clear pattern. The sign of the mean error can be interpreted as follows (in reference to Figure 6.6): when a non-oversampled image is shifted to the right by a quarter of a pixel, the estimated location of a corresponding feature detected in the shifted image is, on average, about -0.15 pixels to the left of where it would ideally be detected.

Lastly, for a half-pixel shift along both x and y directions (as an example of the worst-case error), histograms of the localisation errors (along the x-axis and the Euclidean distance errors) were generated. These histograms are plotted for the cases of no oversampling (Figure 6.7), two-times oversampling (Figure 6.8), and four-times oversampling (Figure 6.9). Although not demonstrated here, shifting along either x or y axis tends to result in comparable localisation errors in both axes.

In general, the feature matching performance exhibits approximate periodicity for shifts outside of the range used in these experiments, as well as symmetry about the y-axis. The trends are reflectionally symmetric for the unsigned metrics (inlier ratio and Euclidean errors) and rotationally symmetric for the signed metrics (estimation bias and signed localisation errors).

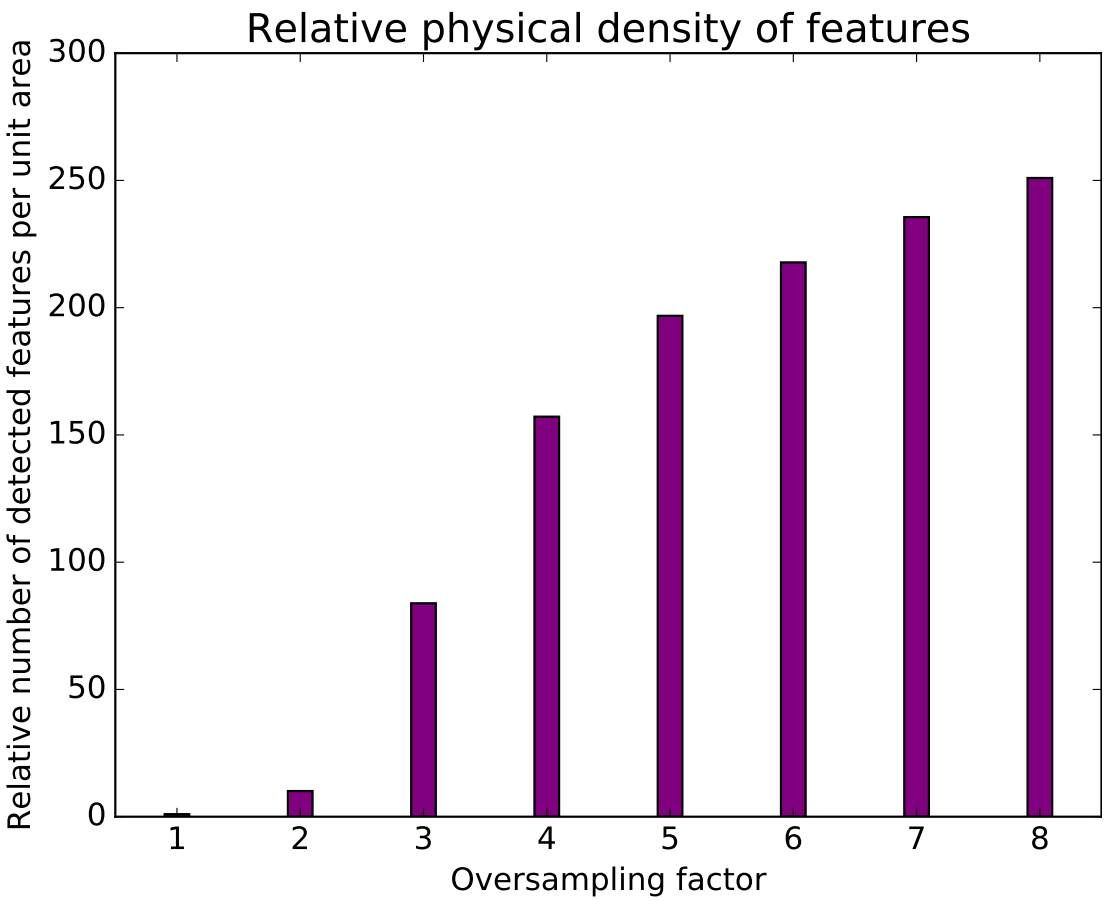


Figure 6.3: The average density of features in terms of physical area, relative to the feature density with non-oversampled images. Oversampling increases the number of detected features significantly, although there is a diminishing return as the oversampling factor increases.

6.3 Discussion

The results shown in Figures 6.2 and 6.3 indicate that 4× oversampling provides the highest number of features, whereas further oversampling does not provide more meaningful detail in the context of SIFT's extrema detection in the Gaussian scale space. Nyquist-sampled speckle images yielded a relatively low number of features, signifying that the resolution of the speckle detail is too low to be reliable (and arguably, repeatable) for scale-invariant feature detection. From an intuitive point of view, it is difficult for humans to recognise (let alone infer) the similarity between the images of Figures 6.1a and 6.1b, even though one in four pixels in the oversampled image share the same corresponding value with a pixel in the original image.

As expected, feature matching performance is stable when an image is shifted by an integer pixel offset, and the observed performance curves are roughly symmetrical about a shift of half a pixel. However, the performance observed with a whole pixel shift is not ideal, having an inlier ratio of approximately 97 % in Figure 6.4. This seems

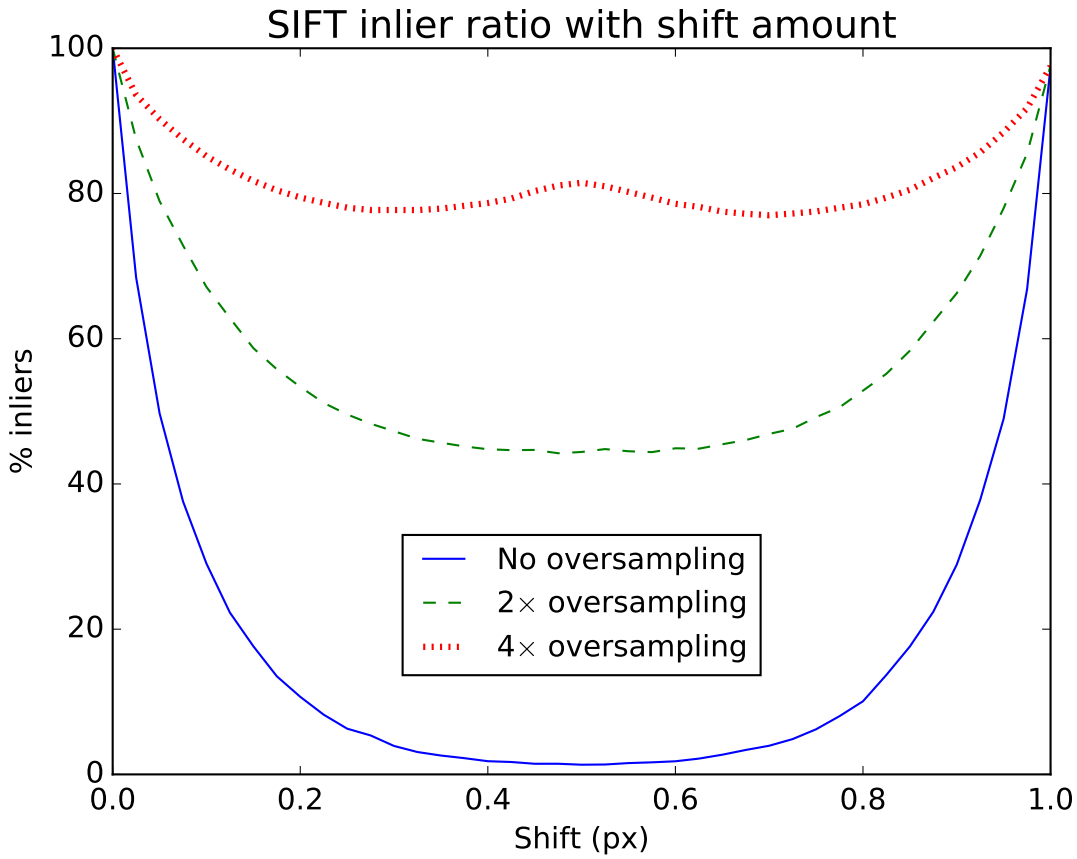


Figure 6.4: The ratio of inlier matches drops quickly as the shift amount approaches half a pixel. This effect is diminished significantly with oversampling.

to be due to the edge effects of the image, with a shifted image not being identical in the Gaussian scale space (even considering the one column of lost information in the shifted image). Further experiments confirmed that the outliers here only occurred near the edges of an image. Anyhow, this effect is of minor consequence and has a lesser impact on larger images.

The image size 100×100 was chosen instead of a larger, more realistic image size in considering the quadratic running time of brute-force feature matching. SIFT performance is not identical across different sizes of images. For example, the average number of detected features (as depicted in Figure 6.2) differs with other image sizes. However, doubling or halving the size of the source images has a relatively minor effect and does not change the general trends observed.

Localisation accuracy degrades with images that are subsampled using a sinc-interpolated sub-pixel shift. On average, the performance of feature detection and matching was observed to be poorest with a half-pixel shift, except in the case of four-times oversampling. When the image consists of uncorrelated speckle noise, i.e., there is no oversampling, the proportion of inliers is prohibitively low (about 1 %) for

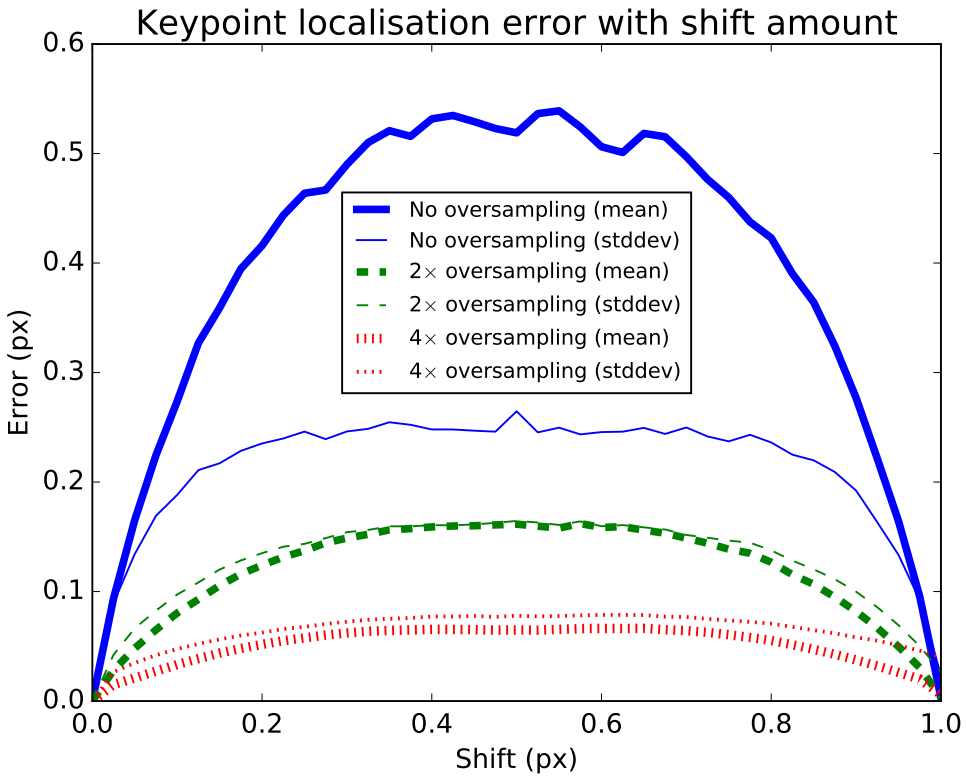


Figure 6.5: The mean localisation error of inlier SIFT matches across subsampled shifted images is close to zero for integer shifts and maximum for half-pixel shifts. The standard deviation of the localisation error shows similar behaviour. With 2× oversampled images, the mean errors are less than half that of the mean errors without oversampling, while the standard deviation is only slightly lower. With 4× oversampling, the mean errors are further halved compared to 2× oversampling.

this worst-case offset. With two-times oversampling, this dip in performance reaches a minimum at about 45 % instead. With four-times oversampling, a half-pixel shift is no longer the worst case, although the lowest inlier ratio observed is about 77 %. As for the mean Euclidean errors in Figure 6.5, the standard deviation of these errors drops slightly with oversampling, whereas the mean localisation errors are halved for each doubling of the sampling rate up to 4× oversampling. The localisation estimation bias (Figure 6.6) also drops significantly with oversampling.

With Nyquist sampling, the presence of the estimation bias and the fact that the mean localisation error exceeds the shift amount points to a discrepancy between the theoretical sinc-interpolated data and the extrema that the SIFT blob detector finds. The mean signed error being near zero at a half-pixel shift can be explained by the subsampled image having values equally weighted by both neighbouring original samples. The bias resembles the form of a sine function and implies that SIFT detection of extrema (which operates in the Gaussian scale space) presumes smoother shapes than

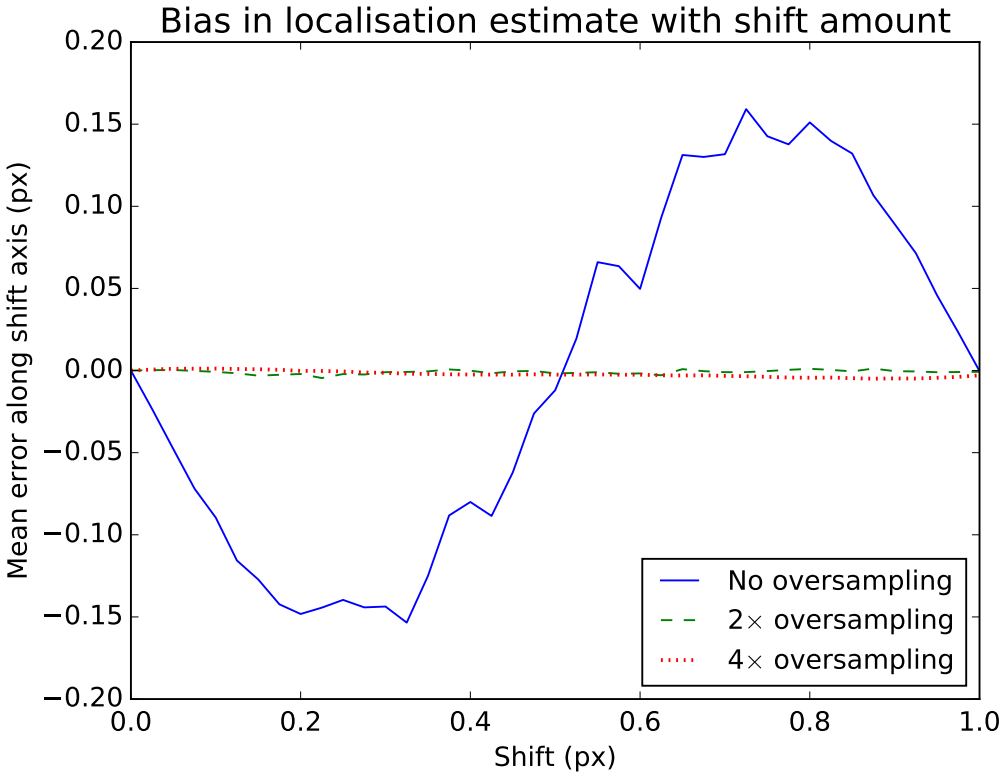


Figure 6.6: With no oversampling, there is a noticeable bias in the detected feature locations with shift amount, resembling a sinusoid. The bias becomes negligible when oversampling is used.

what Nyquist-sampled data predicts. The detector also applies an initial Gaussian blur in order to improve localisation accuracy but assumes that the image is well sampled. Although this blur decreases aliasing, it means that the detected locations do not match up with sinc interpolation.

The error distributions in Figure 6.7 are not particularly recognisable, whereas the shapes of the distributions in Figures 6.8 and 6.9 resemble those from Figures 5.4 and 5.6. As SIFT implementations vary significantly in practice, modelling these specific performance characteristics may not necessarily be of general value for other applications.

Overall, these observed characteristics demonstrate that feature matching may be infeasible with images that are not oversampled due to low repeatability of features, unless the pair of images can somehow be guaranteed to be accurately aligned. Feature matching becomes more feasible with an oversampling of factor of four, with significantly more detected features, a high inlier ratio that is robust to fractional shifts, and less than a quarter of the mean localisation error compared to no oversampling. This has positive implications on estimation accuracy with applications that rely on accurately localised correspondences, such as image co-registration.

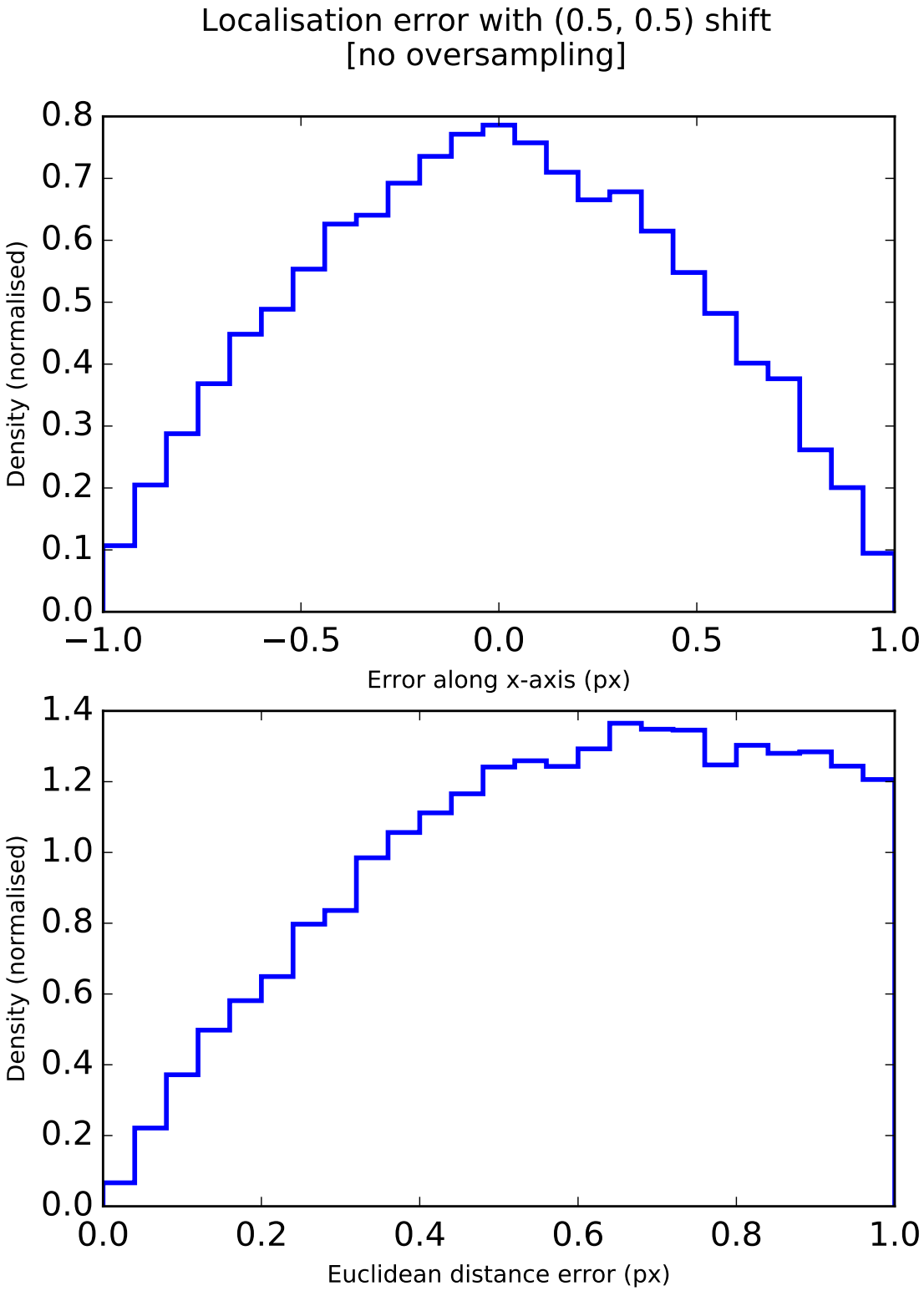


Figure 6.7: Histograms of the signed localisation errors along the x-axis (above) and the Euclidean localisation errors (below) when a half-pixel shift is performed along both directions with no oversampling. The estimated distributions consist of 29,724 inliers (0.48 %) out of 6,156,235 features collected from 200,000 random images.

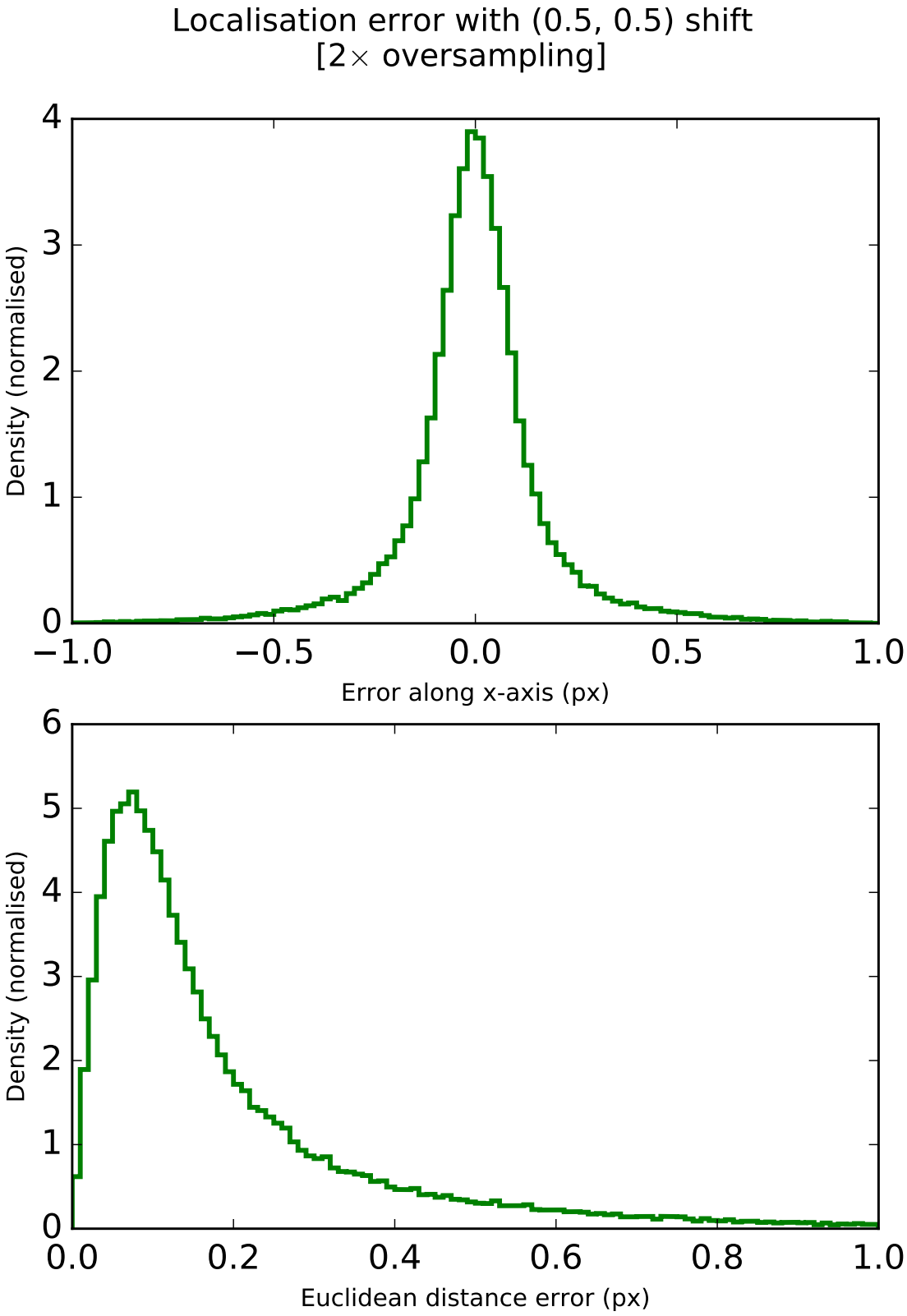


Figure 6.8: Histograms of the signed localisation errors along the x-axis (above) and the Euclidean localisation errors (below) when a half-pixel shift is performed along both directions with 2× oversampling. The estimated distributions consist of 116,930 inliers (37.54 %) out of 311,464 features collected over 4000 random images.

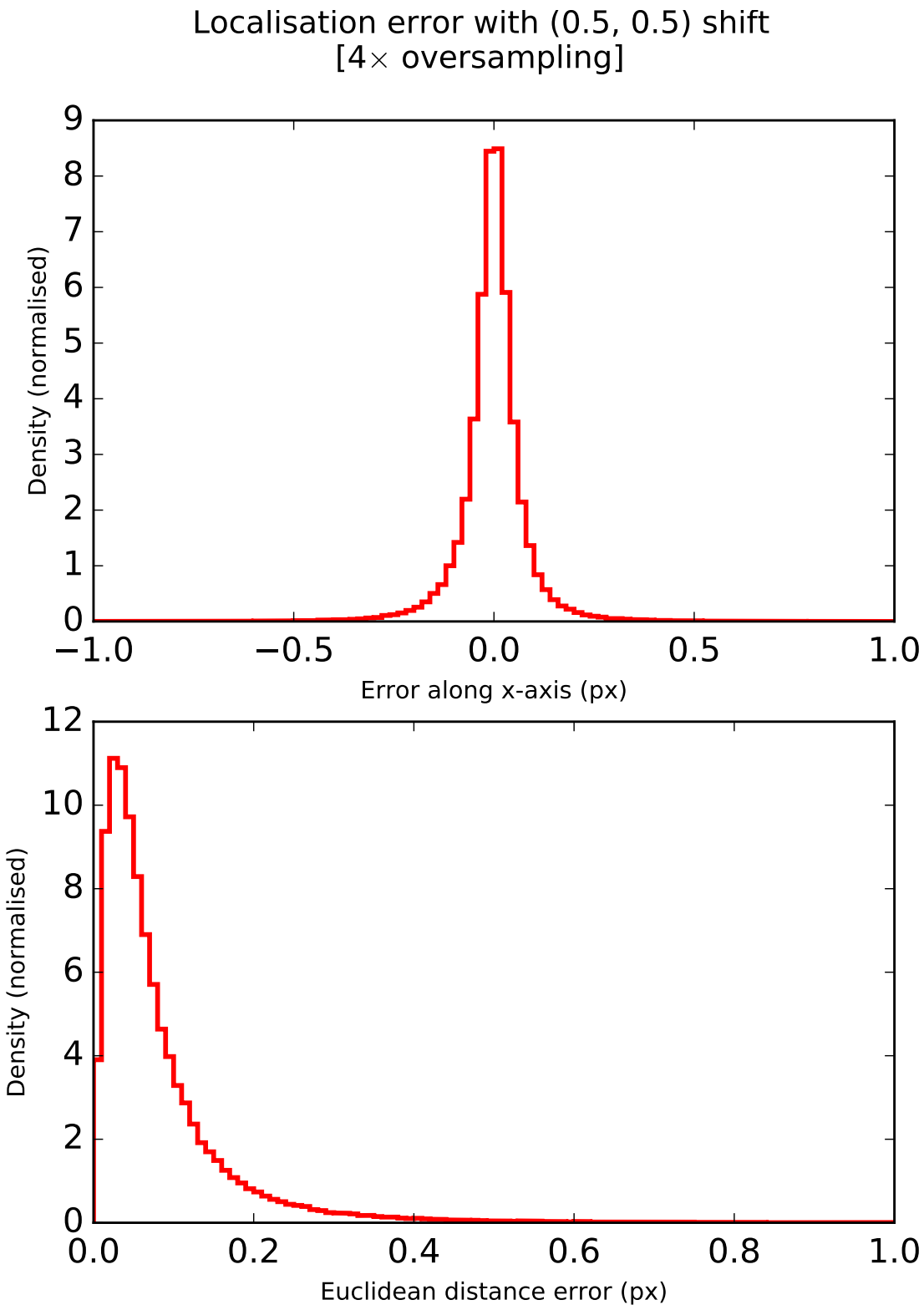


Figure 6.9: Histograms of the signed localisation errors along the x-axis (above) and the Euclidean localisation errors (below) when a half-pixel shift is performed along both directions with 4× oversampling. The estimated distributions consist of 236,008 inliers (76.77 %) out of 307,422 features collected over 1000 random images.

Chapter 7

Feature matching performance on correlated speckle image pairs

As described in Chapter 5, a pair of images from a simulated SAS scene was co-registered using a feature-based pipeline, resulting in an alignment accuracy within the guideline of a tenth of a pixel for interferometric computation and coherent change detection. However, it is problematic to predict whether this level of accuracy can be achieved with other simulated datasets. Real world data also differs significantly from simulated data, making it difficult to characterise how the properties of a sonar system and the imaged scene affect the feasibility of feature matching for coarse initial co-registration. Efforts to procure additional simulated SAS data from the same simulator as in Chapter 5 were unsuccessful. A large dataset is required to meaningfully characterise feature matching performance due to the inherent variability of speckle images, feature matching, and estimation results. Such a dataset must also be simulated in order to have a precise known ground truth.

In general, successful repeat-pass registration hinges on achieving a high coherence between images, even in change detection applications where parts of the scene are expected to be decorrelated. There are several sources of decorrelation, including baseline (spatial) decorrelation, temporal decorrelation, footprint shift (misregistration), and additive acoustic noise. While the coherence between two images for a given sonar system and scene can be modelled as the product of each of these coherence factors, each of these sources of decorrelation contributes to the differences between the co-registered images in different ways. Here, we consider coherence as a single factor, ignoring the finer details of how these sources of decorrelation may interact.

This chapter describes the use of randomly generated images to measure and analyse various aspects of feature matching performance. Section 7.1 describes how pairs of correlated speckle images can be randomly generated. An experimental procedure is outlined in Section 7.2, and feature matching performance on correlated pairs is evaluated for the case of ideal bland scenes. Section 7.3 introduces the concept of feature repeatability and presents the trend for correlated bland images when using SIFT and SURF. Section 7.4 describes an ad-hoc method of mimicking non-bland cor-

related speckle images and presents the feature repeatability for a given underlying ripple scene. Section 7.5 proposes a model for predicting the repeatability of features in relation to scene coherence. This model is shown to provide a reasonable fit to eight different trends: using SIFT or SURF; with or without Lowe's ratio test; with a bland scene or a sand ripple scene. Section 7.6 discusses the overall results and implications.

7.1 Generation of correlated speckle image pairs

Due to the coherent nature of sonar imaging, the measured echo from a rough surface is essentially random yet deterministic. This echo response is an aspect-dependent interference pattern formed by the combined echoes from multiple randomly positioned rough scatterers, where the mean intensity is the desired backscatter coefficient. In the case of SAS, where the resolution size is usually large compared to the sonar wavelength, there are many independent scatterers in each resolution cell. Illuminating these scatterers with a coherent source results in the constructive and deconstructive interference that is speckle, which gives a granular appearance to a sonar image [Fortune et al. 2004]. Speckle is also responsible for a high feature count with feature detectors [Schwind et al. 2010].

Speckle intensity is commonly modelled as a coherent signal-dependent random phenomenon with negative exponential first-order statistics [Bovik 1988][Lee et al. 1994], which is multiplicatively modulated with the scene intensity as follows:

$$I(x, y) = V(x, y)\, U(x, y), \tag{7.1}$$

where $I(x, y)$ is the real-valued intensity at image coordinate (x, y), $V(x, y)$ represents the underlying (intensity) reflectivity of the scene region, and $U(x, y)$ is the speckle intensity, which has a negative exponential distribution with unit mean (and variance). To simplify the analysis, the speckle noise statistics are assumed to be constant over the whole image. Although additive noise is also present in practice, it is ignored in this model.

This multiplicative model represents the case of fully developed speckle, which does not always apply and is only accurate when there is a small change in contrast within each resolution cell. Otherwise, if the spatial details within the cell cannot be resolved by the coherent imaging system, the model is inaccurate [Tur et al. 1982]. The above model is equivalent to modelling the coherent speckle noise as multiplicative circular Gaussian noise with zero mean and unit variance, where the speckle magnitude at each pixel follows a Rayleigh model. To some extent, the statistics of small patches with low contrast can also be approximated as Rayleigh distributed. For larger patches, the distribution is non-Rayleigh and is better modelled using the generalised K-distribution [Lyons et al. 2010][Jakeman and Pusey 1976][Dunlop 1997]. However,

the Rayleigh model remains appropriate for the purposes of simulating speckle magnitude prior to multiplicative modulation with a specified scene image.

A single random speckle image can be generated using the model from (7.1). However, a speckle coherence ρ can also be specified between pairs of randomly generated images using the following method:

Firstly, two matrices G_1 and G_2 of the same shape as the source image are generated with zero-mean unit-variance circular Gaussian noise. The two complex speckle magnitude images are [Bonnett 2017]:

$$A_1(x, y) = G_1(x, y), \tag{7.2}$$
$$A_2(x, y) = \rho^* G_1(x, y) + \sqrt{1 - |\rho|^2}\, G_2(x, y). \tag{7.3}$$

Although the coherence, ρ, is complex valued, note that applying a phase shift does not affect the underlying circular Gaussian statistics in (7.3). Therefore, it is sufficient to specify purely real coherence values. Thus, using the same underlying scene, $V(x, y)$, the pair of correlated complex magnitude images are

$$H_1(x, y) = \sqrt{V(x, y)}\, A_1(x, y), \tag{7.4}$$
$$H_2(x, y) = \sqrt{V(x, y)}\, A_2(x, y), \tag{7.5}$$

and the real-valued intensity images are

$$I_1(x, y) = |H_1(x, y)|^2 = V(x, y)\, U_1(x, y), \tag{7.6}$$
$$I_2(x, y) = |H_2(x, y)|^2 = V(x, y)\, U_2(x, y), \tag{7.7}$$

where the exponentially distributed speckle intensities are

$$U_1 = |A_1|^2, \tag{7.8}$$
$$U_2 = |A_2|^2. \tag{7.9}$$

Note that the magnitude images $|H_1|$ and $|H_2|$ have an expected correlation of ρ, whereas the intensity images I_1 and I_2 have an expected correlation of ρ^2. In practice, the coherence between two images varies throughout the scene due to the sonar geometry for which the baseline varies, as well as due to the other sources of decorrelation. However, it is appropriate to consider a constant coherence for the sake of quantitative analysis of the relationship between coherence and feature matching performance.

As previously described in Section 5.5.1, the images are converted to greyscale prior to feature matching, with a chosen dynamic range of 30 dB. (This is equivalent to using log-intensity images instead while doubling the dynamic range.) The original speckle statistics are not preserved in these clipped log-scale images.

7.2 Feature matching performance of SIFT on correlated bland images

Several experiments were performed in order to investigate various aspects of the performance of features. Randomly generated speckle images of size 200×200 were used for these experiments.

To generate a single statistical sample of a feature matching run, the following steps were involved, regardless of the feature detector/descriptor used:

1. Generate a pair of random correlated speckle images using a specified amplitude scene. Convert these images to greyscale images showing a 30 dB dynamic range.
2. Perform feature detection on each of the two images, obtaining two sets of detected features.
3. Compute feature descriptors for each of these features.
4. Perform brute-force matching, yielding tentative pairs of feature matches.
5. Filter out weaker matches using Lowe's ratio test with a chosen threshold. (This step can be skipped if the threshold is one.)
6. Remove feature matches in such a way as to ensure distinct feature locations within the set of feature matches. (This step is only required for SIFT.)
7. Remove outlier feature matches. A match is an outlier if its two features are located further than one pixel apart (relative to the ground truth of zero offset).

These steps were repeated a number of times with independent pairs of speckle images, with the sample averages used to characterise the results. This section focuses on bland images, where the underlying amplitude scene is a uniform bland scene with unit amplitude and no phase offset. Since the two images in a pair are in perfect alignment, the exact ground truth is known to be a zero offset for each feature correspondence.

With coherence values ranging from 0.8 to 1.0, the relationship between the average number of correct matches and coherence is observed in Figure 7.1, where each data point is the average computed from 100 sample runs. (Lowe's ratio test was not performed.) The maximum number of correct matches occurs at a coherence of one, i.e., when the two images are identical. This value is also equivalent to the average number of detected features (after removing redundant features). Thus, for a random bland speckle image with no oversampling, the OpenCV implementation of SIFT finds an average of 23 distinct features per area of $100\,\text{px} \times 100\,\text{px}$. The average number of inlier matches drops quickly as coherence decreases, with less than five percent of the expected number of potential matches being valid at $\rho = 0.9$.

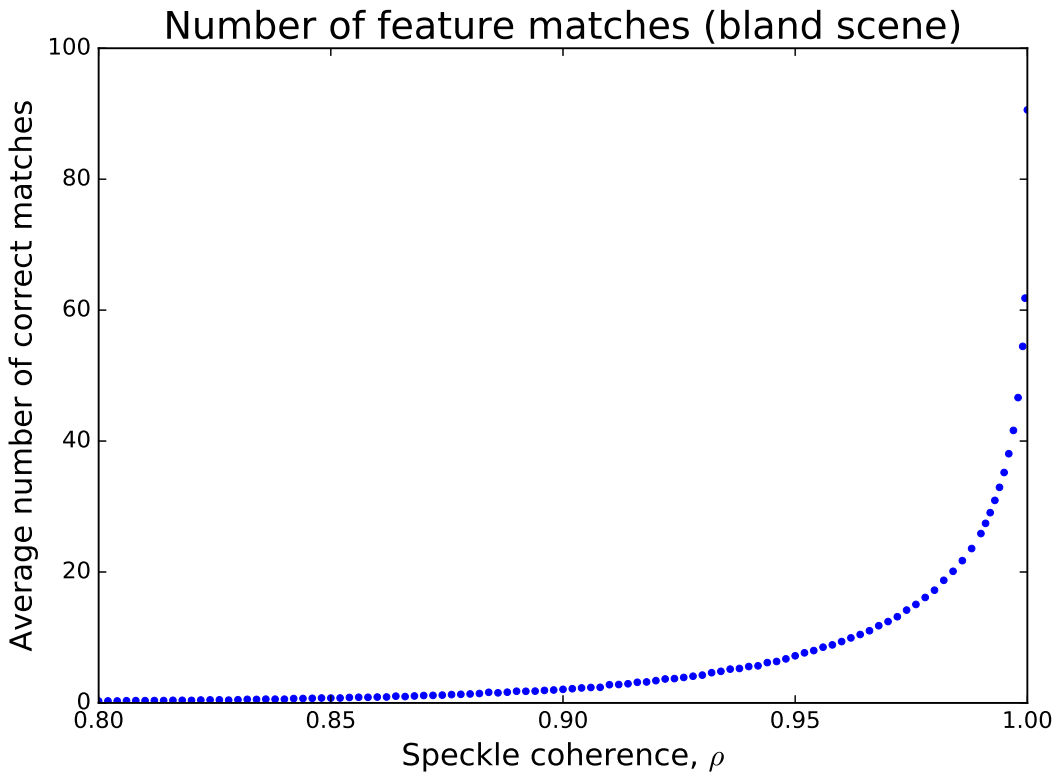


Figure 7.1: The trend between the average number of inlier matches with varying coherence when using SIFT, where the ratio test is not performed. The number of matches decays rapidly as coherence decreases, with almost none being found below 0.8 coherence. The maximum value at 1.0 coherence represents the average number of detected features (91) in an image of size 200×200 pixels.

Next, under the assumption that features found by SIFT are distinctive and independent, the distribution of the localisation error between matched features was estimated for coherence values of 0.97, 0.99, and 0.999. These sample distributions shown in Figure 7.2 are the combinations of 50,000 errors each, taken from many independent runs. In the top plot, the along-track offset distributions appear to be approximately zero-mean and symmetrical about a central peak. Compared to Gaussian distributions, these error distributions have a taller peak but longer side tails and resemble the shape of the distributions in Figures 5.4 and 5.5. The Euclidean distance errors (see bottom plot) are the Euclidean norms of the along-track and across-track error components. The average error increases as coherence decreases. The Euclidean error distributions somewhat resemble Rayleigh distributions but have longer tails and have a similar shape to the distribution in Figure 5.6. Although outlier matches were removed from both these error plots, the majority of the inliers also have less than half a pixel error even for the lowest coherence depicted.

At a coherence of one, the distribution of the number of distinct features (and thus

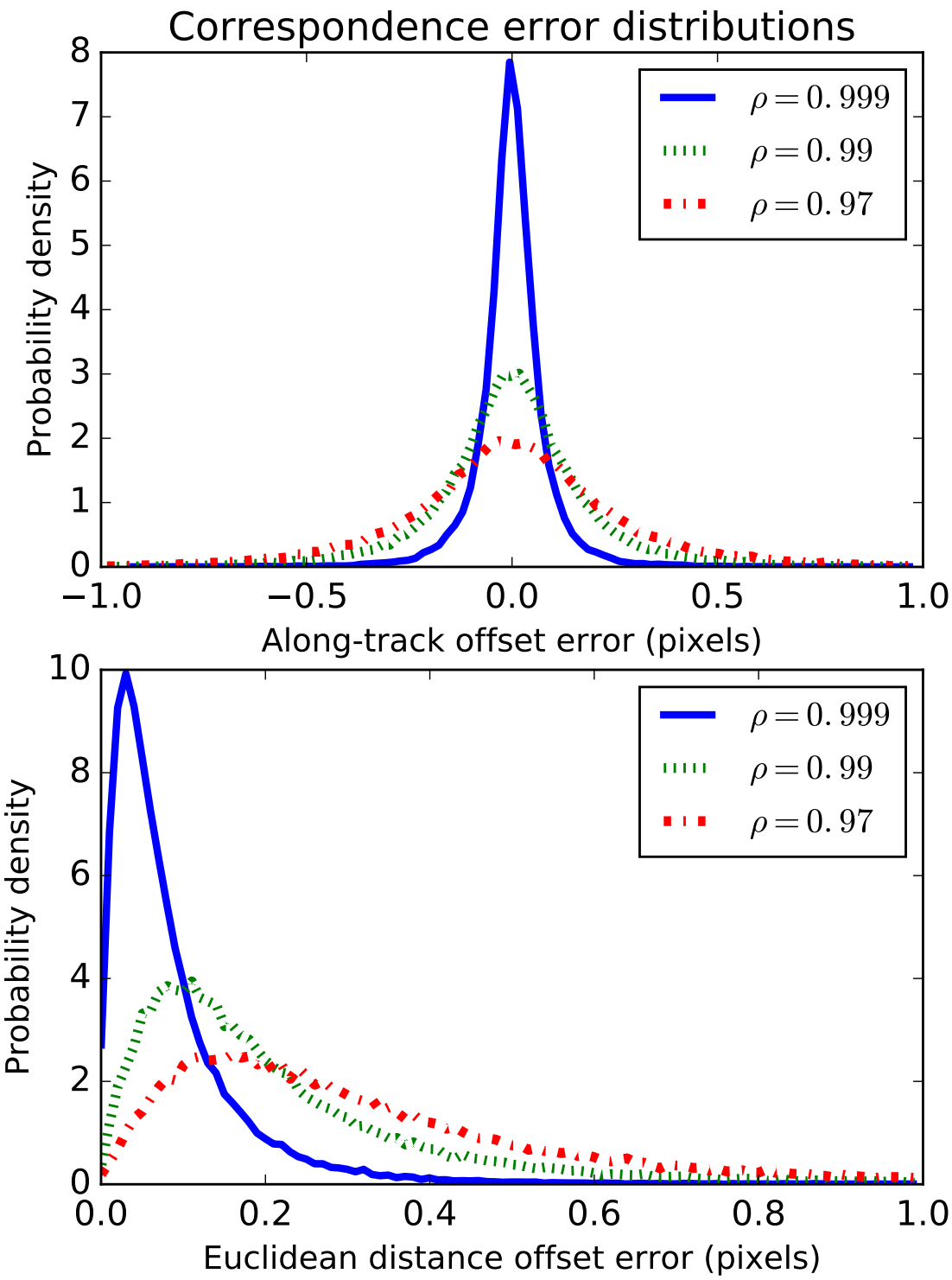


Figure 7.2: The distribution of the localisation errors of SIFT correspondences for the coherence values 0.97, 0.99, and 0.999. The top plot captures the offset errors (in pixels) in the along-track direction. (Theoretically, the across-track errors and along-track errors have an equivalent distribution.) The bottom plot shows the offset errors in terms of their Euclidean distance (in pixels).

feature matches) appears to satisfy the assumptions of a Poisson distribution: features are sparsely distributed throughout the image; features occur independently (for a perfectly bland scene, once redundant features are eliminated); and features have sub-pixel locations. Using the sample mean of 91 matches at coherence one from Figure 7.1, a Chi-squared dispersion test [Brown and Zhao 2002][Cochran 1954] was used to test the fitness of the Poisson model, yielding a p-value of 0.92, which is not inconsistent with the null hypothesis of the underlying distribution being Poisson. Visual comparison with random Poisson samples on a Q–Q plot [Wilk and Gnanadesikan 1968] with 1000 samples also showed a convincing fit. Poisson modelling was also performed for $\rho = 0.9$, where most of the features found in each image do not appear in the other image and so are unmatched. The results were also consistent with a Poisson model, with a p-value of 0.97 found using the dispersion test. While the underlying random process differs from the special case of coherence one, the necessary assumptions for a Poisson distribution still seem plausible as an approximation.

Coherent repeat-pass applications require co-registration to a high degree of accuracy that may not always be achievable using feature-based registration. Although feature-based registration can be used as a preliminary step before refining the initial registration using area-based methods, it is still important to achieve sub-pixel accuracy. In order to do so, feature-based image co-registration relies on both a sufficient number of feature matches and sub-pixel localisation of these matches. As the coherence falls, both of these performance characteristics suffer. As observed in Section 5.7, having an adequate number of matches appears crucial to the stability and reliability of robust estimation. In theory, the localisation error distribution of feature correspondences is also a limiting factor for the resulting accuracy of estimation from these matches. The combination of these factors implies a strong limitation on the practical use of features in relation to the coherence between multi-temporal images. For example, there are virtually no matches below 0.8 coherence, and it also seems unlikely that feature-based registration could be feasible for a pair of bland scene images with less than 0.9 coherence. (In such a non-ideal case, correlation may be viable for registration, but use of the phase data would still likely be impractical.)

Note that these performance characteristics are specific to the OpenCV implementation of SIFT, and thus it can only be directly applied to another application using the same implementation. The next section presents a model for the expected number of correct matches (as in Figure 7.1) that can be used to generalise performance for a given algorithm implementation.

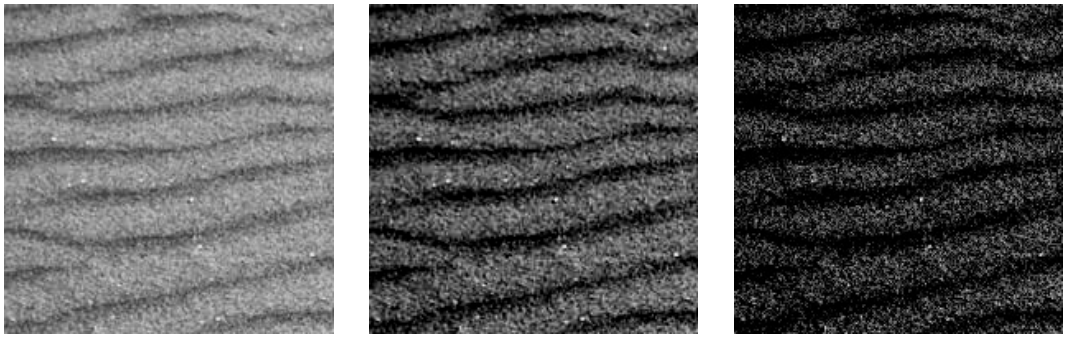
7.3 Feature repeatability of SIFT and SURF for correlated bland images

In Section 7.2, a Poisson distribution was proposed as a suitable model for the number of distinctly located features in an image as well as the number of correct distinct matches between pairs of images with a certain coherence. If the chosen feature descriptor can reliably distinguish features, then matched features can be considered independent from one another. Thus, for a uniform bland scene, it is reasonable to consider that there is a fixed probability that a feature in the reference image is correctly matched to a corresponding feature in the other image. (This probability may conceivably vary based on the number of detected features in each image, but since the two images are not independent (and therefore the expected number of features in the images are not independent), the *a priori* probability can still be considered a constant.) Equivalently, the expected probability represents the likelihood of a local feature remaining recognisable (to the feature detector/descriptor combination) in the other correlated image despite the corrupting influence of speckle noise. In this thesis, this concept is referred to as the "feature repeatability under correlated speckle" or simply feature repeatability. (Note that in the context of feature detectors, the term repeatability can have other meanings or contexts.)

The feature repeatability is directly related to the expected number of feature matches, a trend such as that depicted in Figure 7.1, theoretically differing only by a constant scale factor. Dividing the data points in Figure 7.1 by the average number of detected features forms an estimate for the feature repeatability, although a better estimate is obtained by computing this ratio from total counts over many runs rather than from average values. The feature repeatability trend with coherence was estimated for four different feature matching scenarios on bland scenes: using SIFT for detection and description without Lowe's ratio test, using SIFT with Lowe's ratio test ($r = 0.5$), using SURF without the ratio test, and using SURF with the ratio test ($r = 0.5$). These four estimated trends appear across Figures 7.5 and 7.6 alongside their predicted curves modelled using the method proposed in Section 7.5. Each data point was estimated from 100 runs.

7.4 Feature repeatability for correlated sand ripple scenes

The feature repeatability trend was also estimated for the case of sand ripple scenes instead of bland scenes. As a simple method of emulating speckled sonar images of ripple scenes, random speckle intensity patterns were modulated with intensity images derived from optical images of sand dunes and other ripple-like textures, as in (7.1). Specifically, greyscale optical images were converted to amplitude images by performing exponential scaling, where the highest pixel amplitude value was scaled to 100 (equiv-



(a) Original optical image (b) After contrast adjustment (30 dB log scale) (c) After speckle modulation (30 dB log scale)

Figure 7.3: The 200 px×200 px sand ripple scene used for the results in this chapter.

alent to 40 dB) and the lowest value was scaled to 1 (or 0 dB). Otherwise, the same experimental procedure described in Section 7.2 was followed, but using the modified ripple scene intensity as the underlying amplitude scene.

The ripple scene results in this chapter are based on the optical image in Figure 7.3a, where Figure 7.3b is the image after contrast adjustment and Figure 7.3c shows an example of the scene after random speckle modulation, viewed with a dynamic range of 30 dB. The average density of uniquely located SIFT features was 84 per 100 by 100 pixel area (compared to only 23 with bland scenes). The original image and the exponentially scaled image (before speckle modulation and with 30 dB dynamic range) contained 102 and 128 unique features per 100 px×100 px area respectively. For SURF features, the densities were 140 for the original image, 165 for the contrast-adjusted image, and 205 (on average) after modulation.

The feature repeatability for this scene was estimated using a sample size of 1000 for each coherence value. The trend is depicted in Figure 7.4 alongside its fitted model with estimated parameters $m = 1.46$ and $A = 3.22$. The trend in Figure 7.1 is also shown for comparison. The new model provides a close fit. Four feature repeatability trends for matching with the ripple scene were also modelled, with the results shown across Figures 7.5 and 7.6.

The repeatability trends for ripple scenes based on other images were also estimated, yielding different trends with similarly well-fitted models. For ripple scenes, the expected number of features is consistently higher than for a bland scene, as is the feature repeatability. This pattern is further explored in the following section.

7.5 Model for feature repeatability

The shape of the feature repeatability curves in Figures 7.5 and 7.6 resemble a portion of some form of sigmoid function. Based on experimentation with different models, a

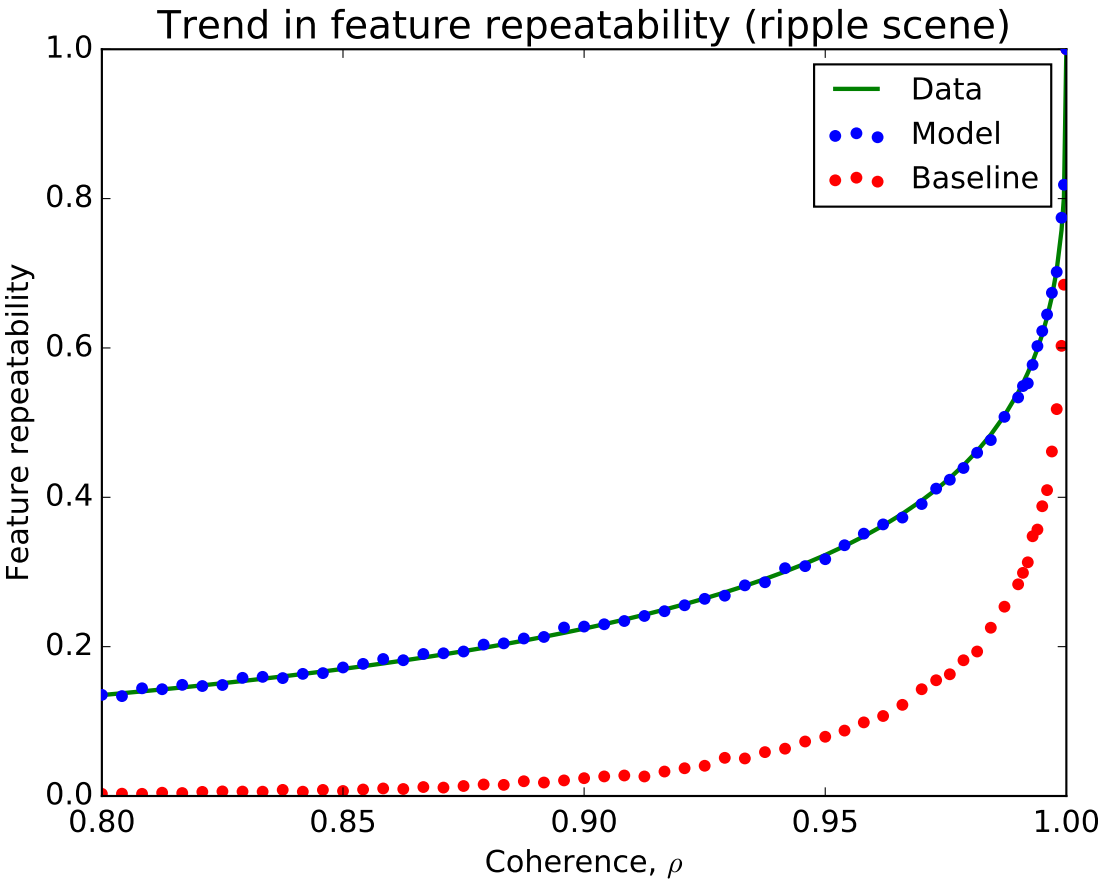


Figure 7.4: Feature repeatability observed from using the sand ripple image in Figure 7.3, fitted with a model using parameters $m = 1.46$ and $A = 3.22$. For comparison, the baseline trend for a bland scene is also shown.

function based on the Gauss error function seemed to give the best and most consistent fit over multiple scenarios. The proposed model for feature repeatability, as a function of coherence ρ and with two constant parameters m and A, is:

$$R(\rho) = 1 - \operatorname{erf}\left(A\,(1-\rho)^{1-\frac{1}{m}}\right). \tag{7.10}$$

The parameters m and A control the shape of the function. Parameter A specifies the feature repeatability at zero coherence (where a larger value means a lower repeatability) and m controls the rate of decay of coherence (smaller value means more rapid decay). These two parameters can be determined exactly given two data points of coherence and repeatability. For example, given two estimated points on the curve,

(ρ_1, R_1) and (ρ_2, R_2), m and A can be calculated as:

$$m = \frac{\ln \frac{1-\rho_1}{1-\rho_2}}{\ln\left(\frac{1-\rho_1}{1-\rho_2}\frac{\text{erf}^{-1}(1-R_2)}{\text{erf}^{-1}(1-R_1)}\right)}, \tag{7.11}$$

$$A = \frac{\text{erf}^{-1}(1-R_1)}{(1-\rho_1)^{1-\frac{1}{m}}}. \tag{7.12}$$

Due to noise in the estimates and the inexactness of the model, it is more suitable to estimate the model using more data points than theoretically required. For simplicity, fitting was performed via a rough initial fit followed by local optimisation to compute a least squares solution.

Repeatability trends were estimated and modelled for the 2×2×2 scenarios: the scene was bland or generated from the scene ripple image in Figure 7.3; SIFT or SURF was used for detection and description; and Lowe's ratio test was performed with threshold $r = 0.5$ or not performed. The four results when using SIFT are shown in Figure 7.5 and the results using SURF are shown in Figure 7.6. The fitted model for each case is also plotted. Each data point is the proportion of retained (inlier) matches estimated over 100 runs. A reasonable model fit is achieved in all cases, with the weakest fit being for a bland scene using SIFT with the ratio test. The corresponding sets of fitted parameters are displayed in Table 7.1. The expected number of unique features is also listed; multiplying the repeatability curves by this number yields the trend in the expected number of feature matches with varying coherence. Therefore, a total of three parameters can be used to characterise this aspect of feature matching performance for a given method or implementation.

Table 7.1: The parameter values for the eight fitted models appearing in Figures 7.5 and 7.6. The feature density, also listed, is the number of unique feature locations (equivalent to the number of unique matches at $\rho = 1$) per 100 px×100 px.

	method	m	A	density of unique features
bland scene	SIFT, r=1.0	1.46	3.22	23
	SIFT, r=0.5	2.31	11.45	
	SURF, r=1.0	1.37	3.95	199
	SURF, r=0.5	1.42	7.68	
ripple scene	SIFT, r=1.0	1.42	1.71	84
	SIFT, r=0.5	1.89	3.14	
	SURF, r=1.0	1.54	1.73	205
	SURF, r=0.5	1.76	3.73	

The feature repeatability model (described in Section 7.5) also fits well for other scenes, from which similar general observations can be drawn. The results using a different underlying optical image (of a beach site) are given in Appendix C along with

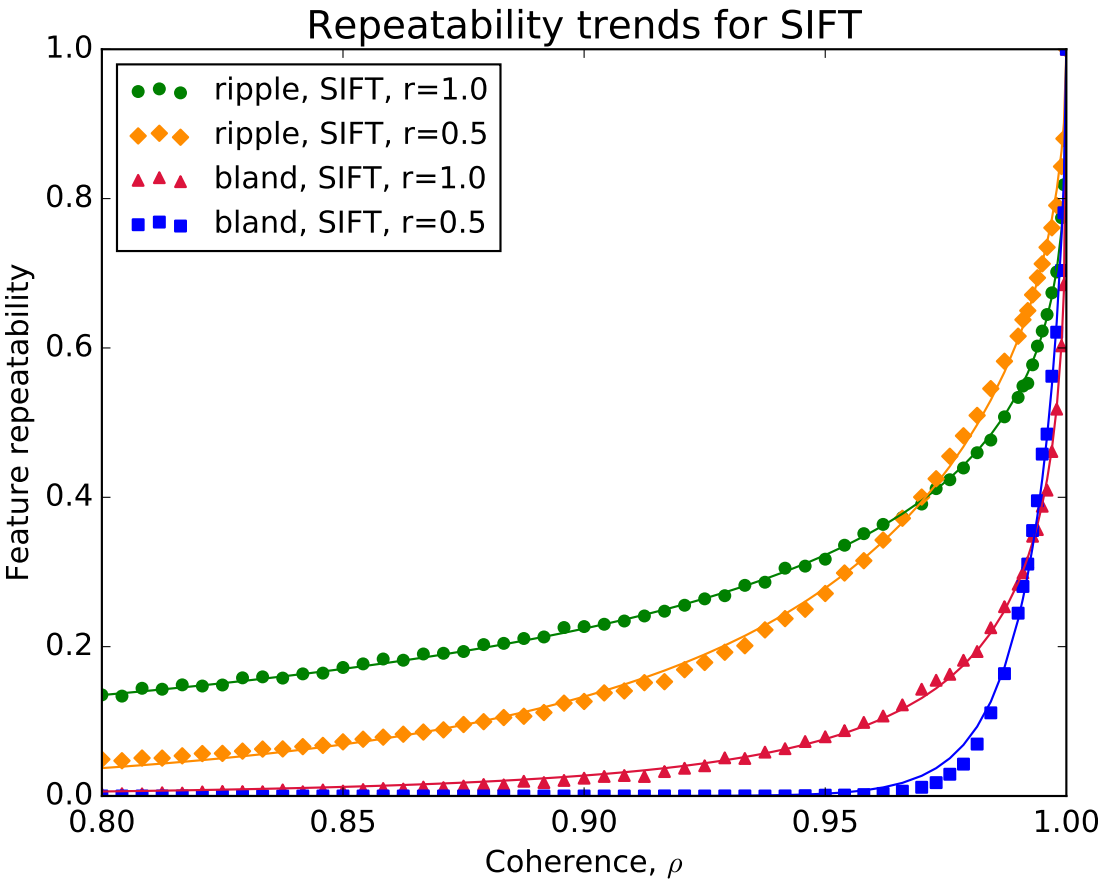


Figure 7.5: Feature repeatability trends using SIFT for both bland and ripple scenes, both with and without performing the ratio test. The trends were estimated from 100 runs. The models fitted to each trend (refer to Table 7.1 for the estimated parameters) are also plotted.

fitted models.

7.6 Discussion

A number of interesting points are observed from the model fitting results. SURF detects significantly more features than SIFT in terms of raw counts. There is a large discrepancy between the number of detected features found by SIFT in relation to scene content, with about 3.7 times more features for the ripple scene than the bland scene, whereas the difference is only slight for SURF. For both detectors, the ripple scene yields a greater repeatability than the bland scene. With the ripple scene used, the speckle modulated images have fewer SIFT features (on average) than both the adjusted optical image and the original image. However, with SURF, the average density of features is roughly the same after speckle modulation. Note that with a non-bland scene there is a loss of contrast due to speckle modulation and clipping to 30 dB dynamic range. The decrease in features for SIFT can be explained by the lower contrast features being

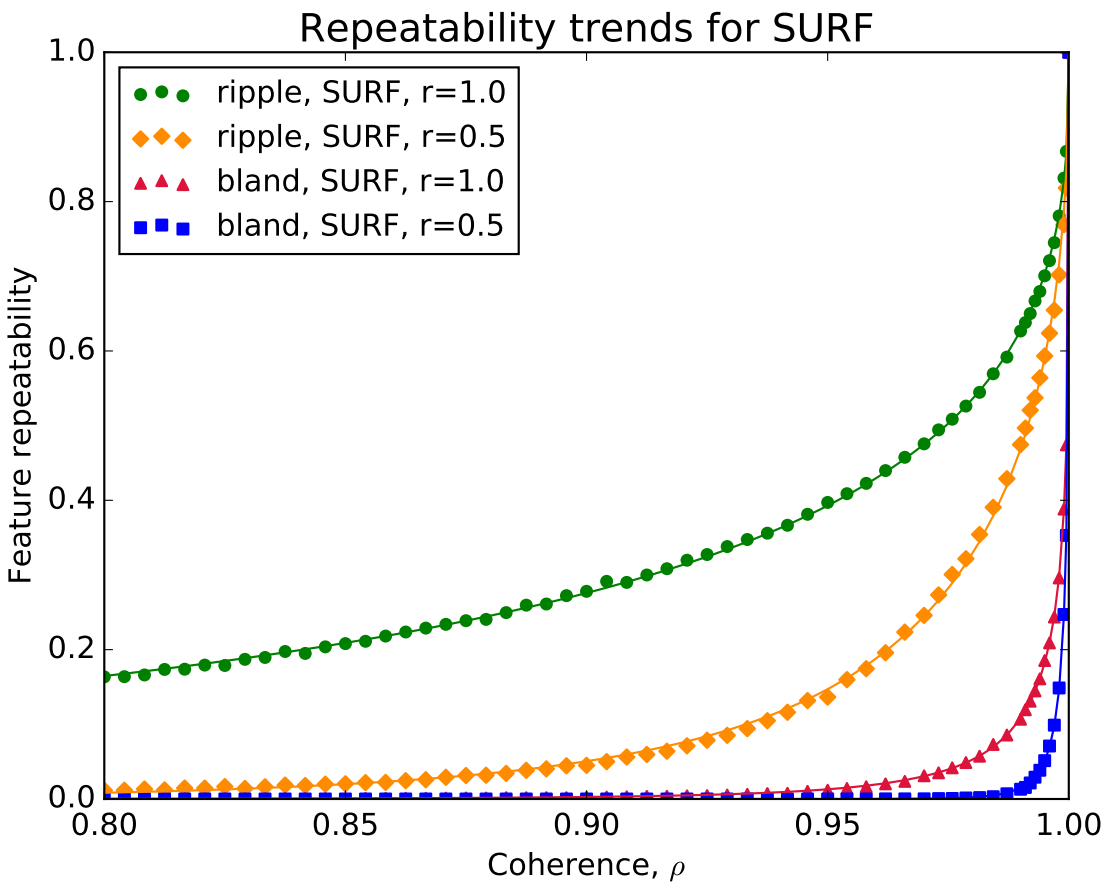


Figure 7.6: Feature repeatability trends using SURF for both bland and ripple scenes, both with and without performing the ratio test. The trends were estimated from 100 runs. The models fitted to each trend (refer to Table 7.1 for the estimated parameters) are also plotted.

suppressed due to the contrast threshold, whereas for SURF the detector sensitivity is based on a Hessian threshold, where a slight decrease in contrast does not strictly imply less distinctive blobs.

A counter-intuitive result in Figure 7.5 is that performing Lowe's ratio test can result in more SIFT inlier matches than not performing it for high coherence values, even though the ratio test only removes matches. (Note that Lowe's ratio test is typically expected to increase the inlier ratio at the cost of number of matches, thus lowering the feature repeatability which is being measured here.) This indicates that there is a noticeable proportion of false matches that the ratio test correctly eliminates as weak matches while preserving good matches. Another possible factor lies in the implementation of the rejection of matches with repeated keypoints, where priority is arbitrarily given to the matches processed first, such that any subsequent matches containing a previously seen feature locations are omitted. For example, if there are three features in the reference image that are matched to the same feature in the second image, then only the first of these pairings is chosen, while the other two are discarded. A improve-

ment would be to choose the match with the smallest distance, which should give equal or better performance on average than when using the ratio test. Nevertheless, the requirement for feature matches to have uniquely located keypoints can increase the impact of the ratio test.

As the coherence drops, the ratio test begins to eliminate a higher proportion of SIFT inlier matches, resulting in lower repeatability (compared to not performing the test) for coherence values below 0.97. This effect is not seen with SURF, where there is a significant drop in repeatability at almost every value of coherence except one, indicating that the ratio test is generally less effective at discriminating weak matches (and preserving good matches) in the SURF descriptor space. Since an adequate number of inlier feature matches are required for stable estimation and there are several RANSAC variants capable of dealing with a low inlier ratio, it seems that Lowe's ratio test is superfluous except in the case where there is an abundance of good feature matches and a portion of these are to be removed to speed up computation.

Overall, the results highlight the idea that while speckle decorrelation degrades the performance of feature matching, speckle noise itself is of potential utility for feature recognition. In the hypothetical case of a perfectly uniform bland scene, feature matching is only possible due to the correlated speckle pattern providing distinguishable image features. Bland scenes yield a lower feature count with SIFT, and feature repeatability degrades severely as the coherence decreases. Ripple scenes contain more robust features overall that have a greater invariance to speckle decorrelation (as seen through the more gradual decay of feature repeatability with respect to coherence), and in this sense the primary contribution seems to be from the scene itself rather than from the speckle pattern. Although not shown, the correspondence error distributions were also found to be smaller overall (in terms of the variances of the along-track and across-track offsets and the Euclidean error), but only to a minor degree. For the ripple scene and without using the ratio test, SURF achieves a higher repeatability on average than SIFT at any value of coherence, whereas SIFT outperforms SURF for the ideal bland scene. With SURF finding over twice as many raw features as SIFT, this means that SURF finds a greater number of inlier feature matches than SIFT in most cases. However, this is not necessarily a point of strength or weakness for either detector, as detector thresholds can be adjusted to find more or less features. Furthermore, having twice as many inlier feature matches at the same level of localisation accuracy has not been shown to reliably contribute to a more accurate registration, while the time taken to perform brute-force matching quadruples.

Based on these results, it is proposed that the repeatability curve for a bland scene acts as a lower bound or baseline expectation of feature matching performance in terms of number of matches (as well as the error distribution, though this is of lesser significance). Although it is possible to construct artificial images where the repeatability would be worse than that for a bland scene, we speculate that natural

images are unlikely to behave this way. Therefore, by evaluating the performance of a specific feature algorithm on bland scenes, it is possible to estimate the number of expected feature matches for an application where there is a known or estimated level of coherence. In practice, this is unlikely to be useful, as the coherence is not typically known until after accurate co-registration. Alternatively, if a specified coherence value corresponds to a repeatability estimate for the "worst-case" scenario of a bland scene, then conversely, a given repeatability estimate corresponds to an optimistic estimate of scene coherence. This implies that it is possible to estimate the level of coherence in a scene by merely performing feature matching on the repeat-pass images of interest and calculating the ratio of correct feature matches to features detected, given a previously estimated repeatability trend for the given feature algorithm. However, due to the relatively high variance of the Poisson distribution, which can be used to model the number of matches for a given coherence, estimates of coherence given a single value of feature repeatability will also have a high degree of uncertainty. It is also difficult to take into account non-uniform decorrelation over the scene due to unpredictable coherence factors such as temporal change.

A major limitation of this experiment is that only speckle decorrelation is considered, whereas other sources of systematic decorrelation do not necessarily follow the same relationship with feature repeatability. For example, the relationship between feature repeatability and decorrelation due to misregistration can be indirectly observed through Figure 6.4, where the shift between 0.0 and 0.5 pixels is related to coherence by (2.37). With all conditions kept the same except for the source of simulated decorrelation, a drop in coherence has a greater effect on feature repeatability in the case of speckle decorrelation. The distribution of feature localisation errors also differs significantly for decorrelation induced by a footprint shift rather than speckle decorrelation. The relationship in the case of additive noise could also be explored.

Although the model fits well to multiple detectors with different parameters and for different scenes, it is purely empirical. Since the model is not an exact fit, the estimates of the model parameters are biased depending on the chosen spread of coherence values for which least squares estimation is performed. When increasing the sample sizes for estimating repeatability, the discrepancy between the model and the data does not disappear; the highest differences between estimated and predicted repeatability tend to be with coherence values close to one (e.g., > 0.99).

Here, the feature repeatability trends for the OpenCV implementations of SIFT and SURF were examined, but the proposed model may be suitable to describe other implementations and algorithms. However, comparing the trends with the ripple scene in this chapter with trends of other ripple scenes and scenes with different textures (such as shown in Appendix C), these trends appear to be dependent on scene content, whereas the repeatability with a bland scene forms the most reliable baseline. Other factors such as detector parameters can also have a large impact on repeatability.

Although rarely discussed in literature, an important choice is the dynamic range of the converted log-magnitude image. As previously noted, increasing the dynamic range can result in a significant increase in number of SIFT features. Regardless of scene content, coherent speckle images tend to have a distinctive distribution of pixel values that may be predictable enough to devise automatic methods of selecting an appropriate dynamic range. For repeat-pass images, a sensible approach is to keep the dynamic range the same across different runs; this could be performed automatically based on analysis of the histogram of magnitude values. As an example, the range could be clipped to the 5th percentile and the 60th percentile of pixel values. Canonically, the largest value in an image is taken as the 0 dB reference for the log scale such that only small values are clamped to the minimum value. However, for scenes consisting mainly of bland textures and only a few bright targets, a low proportion of matched features would neighbour the brightest pixels. Since speckle tends to have long tailed distributions, the occurrence of a singularly bright pixel reduces the useful dynamic range of the rest of the image. Hence, clipping the brightest pixels may be a better strategy. (Additionally, bright spots in one image do not necessarily have the same intensities across repeat-pass images even when the coherence is near one.) The dynamic range should not be too large, otherwise the contrast between low intensity values that are relatively more impacted by aliasing and noise is also preserved. However, these ideas are difficult to test without a large dataset, and an objective performance measure is problematic to construct. It may also be plausible that a new method of converting complex images to greyscale images could result in more repeatable features without resorting to the logarithmic scale.

Chapter 8

Conclusions

A topical application in sonar is image-based change detection, which relies on accurate alignment of repeat-pass images. SAS systems are inevitably subject to imperfect navigation, which leads to artefacts appearing in reconstructed images. Distortion can be minimised by accurately compensating for the navigation data, but this typically requires a combination of state of the art sensors for accurate positioning and data-driven techniques. Registration of speckled imagery is traditionally performed using area-based methods such as correlation, which can require significant amounts of processing time in order to account for various sources of local distortion throughout a synthetic aperture image.

Feature matching (from computer vision) has been adapted for image registration of radar images but remains fairly unused in sonar. Although feature-based methods are less accurate than correlation-based methods, they can significantly reduce computation time using sparse computation, where the relative displacements between two images are specified by a set of feature correspondences rather than being densely sampled throughout the whole scene. Since registration accuracy is of utmost importance for applications requiring coherent processing, the implied use case for feature-based registration is to perform a coarse registration before refining the result using a more accurate correlation-based method. Ideally this would reduce the overall processing time, since correlation-based registration can take orders of magnitude longer and the size of the search space can be decreased significantly by providing a moderately accurate initial estimate of registration. With feature-based registration of speckled images still being in its infancy, the practicality of feature-based SAS registration is still to be understood more clearly.

A feature-based SAS registration method was proposed as a proof of concept. Using an ideal model of a sonar track, it was shown to achieve sub-pixel registration accuracy under ideal conditions. For a high quality simulated SAS scene, a registration to within 0.03 pixels was demonstrated. The registration pipeline is based on the popular combination of feature matching using SIFT features and RANSAC for robust estimation. Based on the literature and preliminary testing, SIFT was chosen

as the most suitable feature detector/descriptor. The novelty of the proposed method is the estimation of the *track registration* (or baseline), which can then be used to compute an image registration. This differs from the usual approach of directly estimating the image registration, since the track registration data can potentially be used to refine navigation data estimates and thus reconstruct improved images. However, the assumptions required for the proposed geometric model are somewhat limiting in practical systems, which are often non-ideal; real imagery may be undersampled, distorted, noisy, or decorrelated. Nevertheless, track registration can be extended to more complicated models such as piece-wise linear models of a sonar track.

Accurate image co-registration is not always possible, especially when the temporal decorrelation between repeat passes is too severe for correlation to give accurate estimates of displacement. Although feature-based registration is presumed to be infeasible in such cases, it is still useful to consider the conditions under which feature-based registration may be viable. To consider more generalisable characteristics of feature matching on speckled imagery, it is necessary to use simulated data, which allows many independent trials to be performed. This would not be possible with real imagery due to inevitable sources of decorrelation. (Also, for SAS in particular, trials at sea would be prohibitively expensive and time-consuming.) Another inherent limitation of real world imagery is the lack of availability of precise ground truth, which accounts for a gap in the research regarding the localisation accuracy of feature detectors.

Bland speckle images were simulated using the multiplicative Gaussian noise model for fully developed speckle. These simulated images were used to explore the effect of sinc interpolation and oversampling on feature matching performance when using SIFT features. An oversampling factor of four was found to yield the greatest number of features for a fixed image size, also providing robustness against fractional image shifts in terms of the proportion of inlier matches and keypoint localisation accuracy. With non-oversampled speckle images, the expected inlier ratio drops below 1 % when a source image is subsampled with a half-pixel offset along both axes. This dip in feature repeatability is greatly alleviated with oversampled images, and the mean localisation error of the inlier feature matches decreases by a factor of two when increasing the sampling rate from 1× to 2× or from 2× to 4×. A bias in the localisation of SIFT keypoints was observed in relation to sub-pixel shift, showing that SIFT feature detection is not concordant with sinc interpolation. However, sampling at twice the Nyquist frequency diminished this effect significantly.

To explore the general feasibility of feature-based registration, the feature matching performance of SIFT and SURF was evaluated in relation to decorrelation of the speckle pattern and the scene content. Local speckle patterns are shown to provide a basis for feature matching of bland images, although non-bland scene content also contributes significantly to the robustness of features. When speckle coherence is relatively low, such as below 0.8, feature matching is shown to be problematic (especially

for bland scenes) due to both low feature repeatability and localisation errors being distributed with greater variance. Although the combination of feature matching performance measures such as density of features and mean localisation error are not reliable predictors of feature-based registration accuracy, a degradation in one or both of these factors is likely to have harsh implications on the feasibility of achieving the gold standard of a tenth of a pixel registration accuracy.

8.1 Ideas for further work

Further experimentation with more diverse datasets

There are several aspects of feature matching performance that could be further explored. For example, images where rotation is present, scenes with man-made objects, alternative models of speckle, perspective effects due to the sonar imaging geometry, additive noise, and multi-look images could be considered in terms of their effect on feature matching performance.

Track registration for a piece-wise linear sonar track model

The track registration method proposed in Chapter 5 can be extended to non-ideal tracks. One possible way is to perform ideal track registration for one local strip of the scene at a time to estimate the path at a local track segment. Special attention is required for the problem of merging the resulting piece-wise tracks in a stable manner, especially when the reconstructed images are distorted to begin with due to inaccurate motion estimation/compensation.

Registration using refined correspondences

While feature-based registration offers the advantage of sparse computation, the sub-pixel localisation accuracy of feature detectors is still outclassed by correlation-based methods, which could be used to refine the localisation of feature correspondences. The relationship between localisation accuracy and registration accuracy is yet to be clarified, especially with the less stable estimation geometry of sonar imaging.

Mapping an image or track registration to correlation-based models

If feature-based registration is to be used as a preliminary step before correlation-based registration, a suitable conversion between geometric models must be performed so that accuracy is not lost, otherwise the speed advantage of incorporating feature-based registration may be negligible. What degree of accuracy can be retained?

Feature matching on 2.5D images

With repeat-pass InSAS imagery, detectors such as SIFT can be adapted to three dimensions, as has been done for volumetric ultrasound [Ni et al. 2008]. Could feature matching be improved using bathymetric data? Can feature matching be used to detect changes in topography?

Conversion of complex images to greyscale

The canonical method of converting complex images to greyscale images by taking the log of magnitude and clipping to a minimum value can potentially be improved upon by clipping to a suitable maximum value as well. These clipping ranges could be automatically determined from histograms of the pixel magnitudes based on the statistics of speckle. Alternative representations that do not use a logarithmic scale can also be considered for maximising the robustness of the resulting features. A large dataset with a variety of images where feature matching performance can be objectively measured is needed to evaluate one conversion method over another.

Appendix A

Least squares with both scene depths known

For the imaging geometry specified in Section 5.1, consider the hypothetical scenario where the depths H and H' of both scenes (and thus t_z) are known exactly. Given a set of N image correspondences $\mathbf{r}_k \leftrightarrow \mathbf{r}'_k$, $k \in 1..N$, any pair of matched coordinates (5.5) and (5.6) are theoretically related by (5.8), where the values x_k, y_k, x'_k (usually unknown when H is unknown), and y'_k are known. The track registration problem can be solved by choosing two system equations based on (5.10):

$$x = x' \cos\alpha + y' \sin\alpha + t_x, \tag{A.1a}$$

$$y = -x' \sin\alpha + y' \cos\alpha + t_y. \tag{A.1b}$$

Note that two equations could alternatively have been taken from (5.11). To solve for α, t_x and t_y the equations are linearised using Taylor polynomials for $\sin\alpha$ and $\cos\alpha$. The 3rd order Maclaurin polynomials are excellent approximations within the region $-45° < \alpha < 45°$, and the higher order terms instead use an initial estimate of the angle, $\hat{\alpha}$:

$$\sin\alpha \approx \alpha - \frac{\hat{\alpha}^3}{6}$$

$$\cos\alpha \approx 1 - \frac{\hat{\alpha}^2}{2}$$

(A.1a) and (A.1b) are thus linearised respectively to:

$$x - x' \cos\hat{\alpha} + y' \frac{\hat{\alpha}^3}{6} = y'\alpha + t_x \tag{A.2a}$$

$$y - y' \cos\hat{\alpha} - x' \frac{\hat{\alpha}^3}{6} = -x'\alpha + t_y \tag{A.2b}$$

These equations now transfer to a linear regression problem. In the case of $n > 2$ correspondences, the following system is overdetermined:

$$\begin{pmatrix} x_1 - x'_1 \cos\hat{\alpha} + y'_1 \frac{\hat{\alpha}^3}{6} \\ \vdots \\ x_n - x'_n \cos\hat{\alpha} + y'_n \frac{\hat{\alpha}^3}{6} \\ y_1 - y'_1 \cos\hat{\alpha} - x'_1 \frac{\hat{\alpha}^3}{6} \\ \vdots \\ y_n - y'_n \cos\hat{\alpha} - x'_n \frac{\hat{\alpha}^3}{6} \end{pmatrix} = \begin{pmatrix} y'_1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ y'_n & 1 & 0 \\ -x_1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ -x_2 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ t_x \\ t_y \end{pmatrix} + \boldsymbol{\epsilon}, \tag{A.3}$$

where $\boldsymbol{\epsilon}$ is the regression error term.

If the error is homoscedastic and normally distributed, ordinary least squares gives the maximum likelihood estimate (MLE). When the rotation is small, the dependent variables are close to the along-track and across-track offsets, so that the errors being minimised are roughly the along-track and across-track errors. In practice, the localisation errors of correspondences are not Gaussian distributed.

Note that ideally the slant-range error would be used instead of the across-track error, in which case the residual sum of squares closely resembles the symmetric transfer error as an approximation. However, linearisation of the slant range requires linearisation of a square root; for the sake of simplicity it is not demonstrated here.

The linearisation process does not incur any cost in accuracy (other than the negligible error of approximating sin and cos) since the regression is performed repeatedly, each time updating the estimate $\hat{\alpha}$ until convergence. This method gives excellent estimates, with the registration error typically being less than 0.01 pixels. Unfortunately, in practice only one scene depth estimate is likely to be usable. Even if both depths are known, any inaccuracy in their difference even on the order of 1 cm can greatly degrade the estimation accuracy. One idea to compensate for this is to discard the part of the image with small slant range (e.g. less than 10 m), where the accuracy of depth estimation has the most impact.

Appendix B

Confidence intervals for estimated parameters from feature matches

The question can be asked: Can a confidence interval be constructed for an overall estimate of a track registration based on a set of feature matches? Firstly, it is appropriate to consider an easier question: Can a confidence interval be constructed for a single parameter such as the along-track offset between two images, estimated from a set of feature matches?

As an example, consider a pair of non-identical SAS images of the same scene such as that of Section 5.5, where there is a fixed along-track offset of zero between the two images. The null hypothesis is that this simulated along-track offset of zero does not imply that the along-track pixel offset for feature matches has a mean tending to zero as the number of matches increases. The alternative hypothesis states that a true along-track offset of zero implies an expected value of zero for the mean along-track offset.

A fundamental problem in statistics is to use a sampling distribution to make inferences concerning the underlying population. It is assumed that the values in the sample are statistically independent, where the sample mean is an unbiased estimator of the population mean. The standard error of the mean (SEM) is the standard deviation of the sample-mean estimate of the population mean. A biased estimator for the SEM is

$$\mathrm{SE}_{\bar{y}} = \frac{s_y}{\sqrt{n}}, \tag{B.1}$$

where

y denotes the sample population of along-track offsets,

s_y is the sample standard deviation, and

n is the size of the sample.

When the distribution of the population is known, the SEM can be used to calculate

a confidence interval for the sample mean. For small sample sizes it is not sufficient to estimate the standard error using the sample standard deviation; a correction must be applied to ensure the confidence interval is large enough. For example, a common rule is to construct a confidence interval using the Student's t-distribution when the sample size is less than 30.

Performing feature matching using SIFT and applying the ratio test with a threshold of 0.75 on the pair of images from Section 5.5, a distribution of the along-track offset was calculated from the along-track pixel offsets between matched feature locations. For an unknown distribution, unknown population variance, and a large sample size, the confidence interval is calculated using a z-score, where the sample mean is treated as a realisation of a normal distribution; the central limit theorem (CLT) applies.

After removal of redundant matches as described in [Rabin et al. 2010], the resulting statistics were:

$$
\begin{aligned}
n &= 4980, \\
\bar{y} &= -0.0058, \\
s_y &= 0.188, \\
\mathrm{SE}_{\bar{y}} &= 0.0027.
\end{aligned}
$$

This gives a 95 % confidence interval for the mean along-track offset: $(-0.0110, -0.0006)$. The true offset of zero lies outside this confidence interval. There were no outliers in the samples.

If the samples are somehow positively correlated, the sample standard deviation forms an underestimate, resulting in a confidence interval that is too narrow. To improve the independence of samples, the set of feature matches was filtered in such a way that no remaining matches had feature locations within 7 pixels of other matched locations. The results were:

$$
\begin{aligned}
n &= 2992, \\
\bar{y} &= -0.0085, \\
s_y &= 0.181, \\
\mathrm{SE}_{\bar{y}} &= 0.0033,
\end{aligned}
$$

yielding a 95 % confidence interval of $(-0.0150, -0.0020)$. The value zero lies outside this confidence interval.

In these two cases, the 95 % confidence interval excluded the true value of the along-track offset. It appears that the necessary conditions or assumptions for this statistical procedure to give meaningful results were not satisfied. There are many

possible explanations for this result.

The primary argument for the null hypothesis is that the images are not random, therefore correspondences corrupted by speckle noise are not random, and thus the CLT does not apply. Firstly, the underlying scene in the simulated dataset is predefined. The image obtained depends on the sonar track, including the deterministic speckle pattern observed. Although speckle is random-like, the speckle pattern is correlated across images with similar viewpoints. Furthermore, the effect of speckle interference on feature detection is not independent of the features in the underlying scene. The robustness of a given feature is partially related to its size, magnitude, and orientation. For example, SIFT keypoints with larger scale tend to be more repeatable [Schwind et al. 2010]. Clearly, along-track pixel offsets are not identically distributed even when the values are independent. This invalidates the estimation of the standard deviation, since there is no underlying random variable.

Although the mean pixel offset may be considered as the sum of many independent random variables, the classical CLT requires the variables to be identically distributed. The Lyapunov CLT does not require the sum to be of identically distributed variables, but the speed of convergence to the asymptotic normal distribution is slow in general. A bound can be quantified using Stein's method only if the distributions are known. A confidence interval cannot be constructed, since the variances of the pixel offsets (if they exist) are unknown.

Lastly, the statistical test above is in favour of the null hypothesis; there is no evidence to suggest that the mean along-track pixel offset of feature matches has an expected value of zero. On the contrary, the mean pixel offset may be dependent on the scene, whether or not the observed speckle noise can be considered to be governed by a random field.

In summary, it may be concluded that either the required assumptions for computing a confidence interval are not met, and/or the mean along-track offset estimated from feature matches is not expected to converge to the true along-track offset due to the deterministic nature of the imaging. It may be argued that the mean of means can be expected to tend to the true along-track offset over multiple independent speckle realisations, but this does not reflect the reality of real scenes either. Even if construction of a confidence interval is not possible, a mean offset may still nominally be the best estimate for the true offset.

Can a confidence interval be constructed for an overall estimate of registration? Returning to this question, it has been argued statistically and heuristically that it is infeasible to construct a confidence interval for a single parameter. By extension, it is also infeasible to construct an interval for a set of parameters, which is formed as a more complicated combination of data points from feature matches. The feature

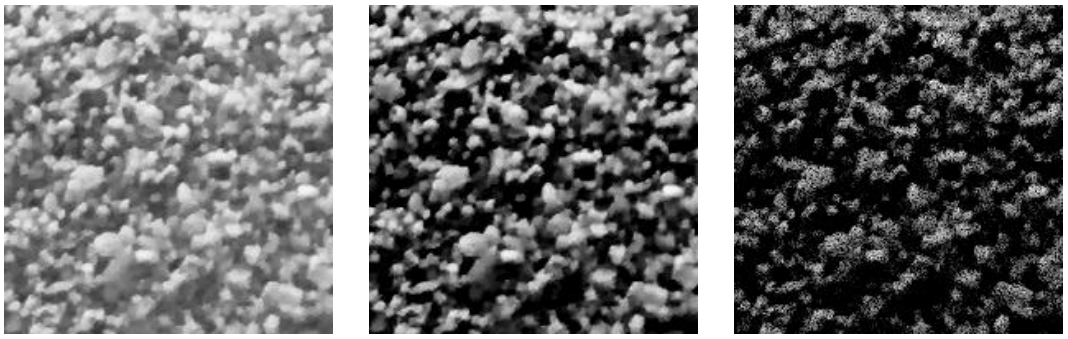
localisation errors do not follow a true statistic, and even if their uncertainty could be modelled, it is still impractical to construct an unbiased estimator at the desired level of accuracy for a confidence interval based on the standard error.

Estimation of track registration parameters is somewhat analogous to the homography estimation problem in computer vision. If there is a suitable model for feature localisation errors (a Gaussian distribution is typically assumed for simplicity), excluding outliers, the best estimate can be specified as the maximum likelihood estimate of the homography. Maximum likelihood estimation cannot be used to construct a confidence interval. The statistical estimation of confidence intervals for multiple parameters is called *bootstrapping* [Davison and Hinkley 1997]. Usually, the underlying distribution of measurement errors is unknown, hence the use of Monte Carlo simulations. Statistical bootstrapping is not considered in homography estimation for multiple reasons: the use of the overspecified 3×3 homography matrix prohibits the use of confidence intervals; the problem of finding the best estimate does not coincide with computing a confidence or uncertainty estimation, and there is rarely a practical need for such statistics. Although bootstrapping differs from the statistical approach to confidence intervals attempted above, it does not offer quantifiable guarantees for finite samples [Efron and Tibshirani 1994], and this particular track estimation problem and dataset may still violate the underlying assumptions required for bootstrapping methods to give satisfactory results.

Appendix C

Feature repeatability for a beach scene

The model for feature repeatability described in Section 7.5 was found to provide a reasonable fit for several images of ripple scenes. As a texturally distinct example of another natural scene, feature repeatability trends for a beach scene are given here. Figure C.1 shows the greyscale optical image at different stages of processing. In the original image, the densities of unique feature locations were 173 for SIFT and 95 for SURF, and the feature densities after contrast adjustment were 183 and 96 respectively. A noticeable difference between this beach scene and the ripple scene in Figure 7.3 is the greater number of SIFT features and lower number of SURF features. While the exact performances of SIFT and SURF are different for this scene, the general observations of the relative performance of each detector made in Section 7.6 still apply.



(a) Original optical image (b) After contrast adjustment (30 dB log scale) (c) After speckle modulation (30 dB log scale)

Figure C.1: The 200 px×200 px image of a beach scene.

Feature repeatability trends and their fitted models are plotted for four scenarios in Figure C.2: using SIFT or SURF, with or without Lowe's ratio test. The model parameters and the expected number of feature matches per 100 px×100 px are listed in Table C.1.

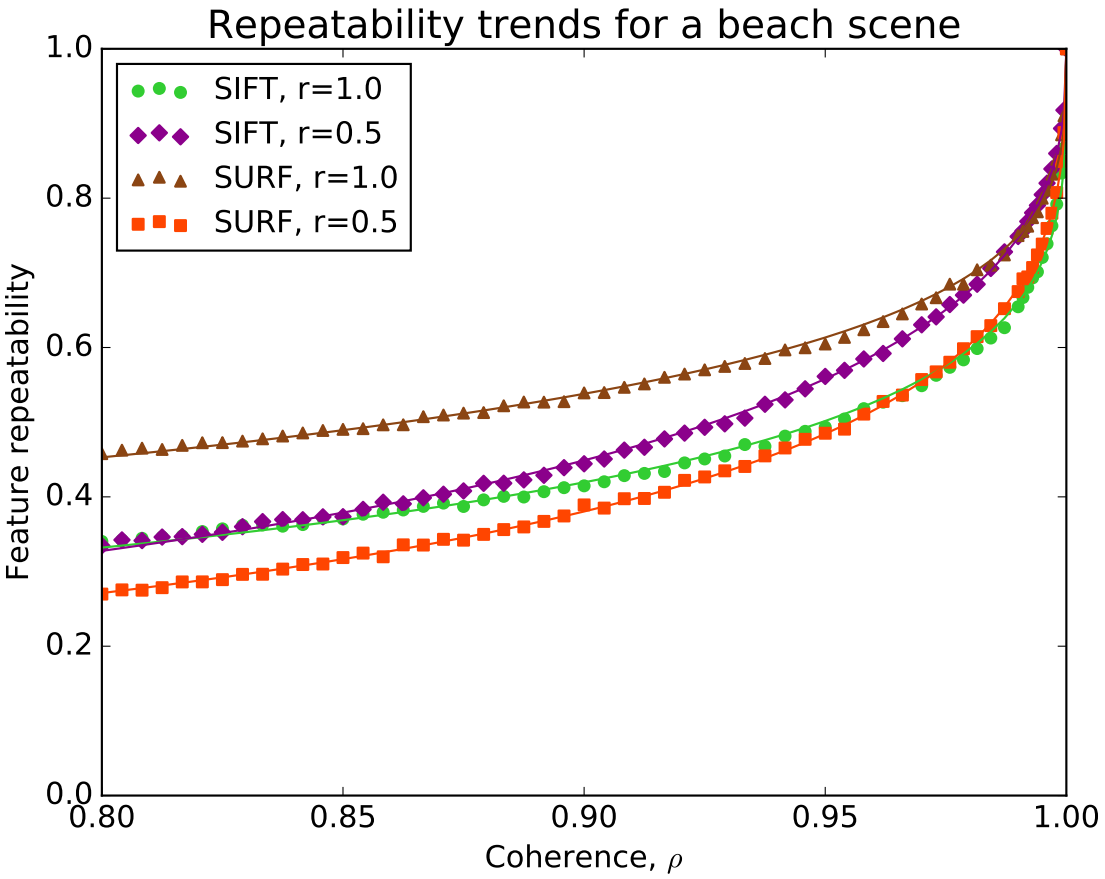


Figure C.2: Feature repeatability trends and their fitted models using SIFT and SURF for a beach scene, with and without performing the ratio test. The trends were estimated from 100 runs.

Table C.1: The parameter values for the four fitted models in Figure C.2 and the density of unique features.

method	m	A	feature density
SIFT, r=1.0	1.36	1.05	129
SIFT, r=0.5	1.59	1.26	
SURF, r=1.0	1.40	0.84	113
SURF, r=0.5	1.49	1.32	

References

Abdel-Hakim, A. E. and Farag, A. A. (2006). CSIFT: A SIFT descriptor with color invariant characteristics. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 1978–1983. IEEE.

Abrahamsen, P. (1997). A review of Gaussian random fields and correlation functions. Technical Report 917, Norwegian Computing Center.

Agrawal, M., Konolige, K., and Blas, M. R. (2008). CenSurE: Center surround extremas for realtime feature detection and matching. In *European Conference on Computer Vision*, pages 102–115. Springer Berlin Heidelberg.

Alahi, A., Ortiz, R., and Vandergheynst, P. (2012). FREAK: Fast retina keypoint. In *Computer Vision and Pattern Recognition (CVPR)*, pages 510–517. IEEE.

Alcantarilla, P. F., Nuevo, J., and Bartoli, A. (2013). Fast explicit diffusion for accelerated features in nonlinear scale spaces. In *Proc. British Machine Vision Conf. (BMVC)*, Bristol, UK.

Andrew, R. K., Howe, B. M., Mercer, J. A., and Dzieciuch, M. A. (2002). Ocean ambient sound: comparing the 1960s with the 1990s for a receiver off the California coast. *Acoustics Research Letters Online*, 3(2):65–70.

Arandjelovic, R. and Zisserman, A. (2012). Three things everyone should know to improve object retrieval. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2911–2918. IEEE.

Au, W. W. L. and Banks, K. (1998). The acoustics of the snapping shrimp Synalpheus parneomeris in Kaneohe Bay. *The Journal of the Acoustical Society of America*, 103(1):41–47.

Azaria, M. and Hertz, D. (1984). Time delay estimation by generalized cross correlation methods. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):280–285.

Barclay, P. J. (2006). *Interferometric synthetic aperture sonar design and performance*. PhD thesis, University of Canterbury. Electrical and Computer Engineering.

Baumberg, A. (2000). Reliable feature matching across widely separated views. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 774–781. IEEE.

Bay, H., Ess, A., Tuytelaars, T., and van Gool, L. (2008). Speeded-up robust features

(SURF). *Computer Vision and Image Understanding*, 110(3):346–359.

Bay, H., Tuytelaars, T., and van Gool, L. (2006). SURF: Speeded up robust features. In *European Conference on Computer Vision*, pages 404–417. Springer.

Bellec, R., Legris, M., Khenchaf, A., Amate, M., and Hetet, A. (2005). Repeat-track SAS interferometry: Feasibility study. In *OCEANS*, pages 748–754. MTS/IEEE.

Bellettini, A. and Pinto, M. (2009). Design and experimental results of a 300-kHz synthetic aperture sonar optimized for shallow-water operations. *IEEE Journal of Oceanic Engineering*, 34(3):285–293.

Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. In *Noise Reduction in Speech Processing*, chapter 5. Springer.

Bentoutou, Y., Taleb, N., Kpalma, K., and Ronsin, J. (2005). An automatic image registration for applications in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 43(9):2127–2137.

Birkhoff, G. D. (1931). Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences*, 17(12):656–660.

Boas, D. A. and Dunn, A. K. (2010). Laser speckle contrast imaging in biomedical optics. *Journal of Biomedical Optics*, 15(1):011109/1–12.

Boker, S. M., Rotondo, J. L., Xu, M., and King, K. (2002). Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological Methods*, 7(3):338–355.

Bonnett, B. (2017). *Coherent change detection in repeat-pass synthetic aperture sonar*. PhD thesis, University of Canterbury.

Bonnett, B., Hayes, M., and Hunter, A. (2013). Registration of images from a hull mounted, low frequency synthetic aperture sonar. In *Image and Vision Computing New Zealand*, pages 142–147. IEEE.

Bookstein, F. L. (1989). Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585.

Born, M. and Wolf, E. (1999). *Principles of Optics*. Cambridge University Press, 7th edition.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Bovik, A. C. (1988). On detecting edges in speckle imagery. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(10):1618–1627.

Bradski, G. (2000). The OpenCV library. *Dr. Dobb's Journal of Software Tools*, 25(11):120–126.

Brown, L. D. and Zhao, L. H. (2002). A test for the Poisson distribution. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 611–625.

Brown, L. G. (1992). A survey of image registration techniques. *ACM Computing Surveys*, 24(4):325–376.

Brown, M. and Lowe, D. G. (2002). Invariant features from interest point groups. In *Proc. British Machine Vision Conf. (BMVC)*, pages 253–262, Cardiff, UK.

Bruzzone, L. and Prieto, D. F. (2000). Automatic analysis of the difference image for unsupervised change detection. *IEEE Transactions on Geoscience and Remote sensing*, 38(3):1171–1182.

Byrne, R. H., Duxbury, A. C., and Mackenzie, F. T. (2017). Seawater.

Cafforio, C., Prati, C., and Rocca, F. (1991). Full resolution focusing of SEASAT SAR images in the frequency-wave number domain. *International Journal of Remote Sensing*, 12(3):491–510.

Calder, B. R. and Mayer, L. A. (2003). Automatic processing of high-rate, high-density multibeam echosounder data. *Geochemistry, Geophysics, Geosystems*, 4(6). 1048.

Callow, H. J. (2003). *Signal processing for synthetic aperture sonar image enhancement.* PhD thesis, University of Canterbury.

Calonder, M., Lepetit, V., Strecha, C., and Fua, P. (2010). BRIEF: Binary robust independent elementary features. In *European Conference on Computer Vision*, pages 778–792. Springer.

Capel, D. P. (2005). An effective bail-out test for RANSAC consensus scoring. In *Proc. British Machine Vision Conf. (BMVC).*

Caporale, S. and Petillot, Y. R. (2017). A new framework for synthetic aperture sonar micronavigation. *Computing Research Repository (CoRR)*, abs/1707.08488.

Caprais, P. and Guyonic, S. (1997). Squint and forward looking synthetic aperture sonar. In *OCEANS*, volume 2, pages 809–814. MTS/IEEE.

Cavicchi, T. J. (1992). DFT time-domain interpolation. *IEE Proceedings F (Radar and Signal Processing)*, 139(3):207–211.

Chang, M. (2006). *Forest Hydrology: An Introduction to Water and Forests.* Taylor & Francis, second edition.

Chum, O. and Matas, J. (2005). Matching with PROSAC – progressive sample consensus. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 220–226. IEEE.

Chum, O. and Matas, J. (2008). Optimal randomized RANSAC. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1472–1482.

Chum, O., Matas, J., and Kittler, J. (2003). Locally optimized RANSAC. In *Pattern Recognition*, pages 236–243. Springer.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, 10(4):417–451.

Cook, C. E. and Bernfeld, M. (1967). *Radar signals: An introduction to theory and application.* Electrical Science Series. Academic Press.

Cover, T. M. and Thomas, J. A. (1991). *Elements of information theory.* John Wiley & Sons.

Crow, F. C. (1984). Summed-area tables for texture mapping. *ACM SIGGRAPH*

Computer Graphics, 18(3):207–212.

da Silva, S. R. B. O. (2009). *Interferometric synthetic aperture sonar system supported by satellite*. PhD thesis, Universidade do Porto (Portugal).

Dainty, J. C. (1980). An introduction to Gaussian speckle. In *Proceedings of SPIE*, volume 243.

Dainty, J. C. (1984). *Laser speckle and related phenomena*. Topics in Applied Physics. Springer-Verlag, second edition.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

de Moustier, C. (1988). State of the art in swath bathymetry survey systems. *International Hydrographic Review*, 65(2).

Dekker, R. J. (2003). Texture analysis and classification of ERS SAR images for map updating of urban areas in the Netherlands. *IEEE Transactions on Geoscience and Remote Sensing*, 41(9):1950–1958.

Dellinger, F., Delon, J., Gousseau, Y., Michel, J., and Tupin, F. (2012). SAR-SIFT: A SIFT-like algorithm for applications on SAR images. In *International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3478–3481. IEEE.

Dellinger, F., Delon, J., Gousseau, Y., Michel, J., and Tupin, F. (2015). SAR-SIFT: A SIFT-like algorithm for SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 53(1):453–466.

Denbigh, P. N. (1989). Swath bathymetry: Principles of operation and an analysis of errors. *IEEE Journal of Oceanic Engineering*, 14(4):289–298.

Dillon, J. (2013). Seeing with sound: Why sonar resolution matters for seabed mapping. `http://www.krakensonar.com/index.php/en/investors/news/34-august-2013-ocean-news-technology-article`.

Dillon, J. and Myers, V. (2014a). Baseline estimation for repeat-pass interferometric synthetic aperture sonar. In *European Conference on Synthetic Aperture Radar (EUSAR)*. VDE.

Dillon, J. and Myers, V. (2014b). Coherence estimation for repeat-pass interferometry. In *OCEANS*. MTS/IEEE.

Douglas, B. L. and Lee, H. (1992). Synthetic aperture active sonar imaging. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 37–40. IEEE.

Douglas, B. L. and Lee, H. (1993). Synthetic-aperture sonar imaging with a multiple-element receiver array. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 445–448. IEEE.

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In *Constructive Theory of Functions of Several Variables*, pages 85–100, Berlin, Heidelberg. Springer.

Duits, R., Florack, L., De Graaf, J., and ter Haar Romeny, B. (2004). On the axioms of scale space theory. *Journal of Mathematical Imaging and Vision*, 20(3):267–298.

Dunlop, J. (1997). Statistical modelling of sidescan sonar images. In *OCEANS*, volume 1, pages 33–38. IEEE.

Dunn, A. K. (2012). Laser speckle contrast imaging of cerebral blood flow. *Annals of Biomedical Engineering*, 40(2):367–377.

Durrant-Whyte, H. and Bailey, T. (2006). Simultaneous localization and mapping: Part I. *IEEE Robotics & Automation Magazine*, 13(2):99–110.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Faccia, P. A., Pardini, O. R., Amalvy, J. I., Cap, N., Grumel, E. E., Arizaga, R., and Trivi, M. (2009). Differentiation of the drying time of paints by dynamic speckle interferometry. *Progress in Organic Coatings*, 64(4):350–355.

Fahy, F. and Walker, J. (1998). *Fundamentals of Noise and Vibration*. Taylor & Francis.

Fallon, M. F., Folkesson, J., McClelland, H., and Leonard, J. J. (2013). Relocating underwater features autonomously using sonar-based SLAM. *IEEE Journal of Oceanic Engineering*, 38(3):500–513.

Fan, B., Huo, C., Pan, C., and Kong, Q. (2013). Registration of optical and SAR satellite images by exploring the spatial relationship of the improved SIFT. *IEEE Geoscience and Remote Sensing Letters*, 10(4):657–661.

Fandos, R. (2012). *ADAC system design and its application to mine hunting using SAS imagery*. PhD thesis, Technische Universität.

Fandos, R., Debes, C., and Zoubir, A. M. (2013). Resampling methods for quality assessment of classifier performance and optimal number of features. *Signal Processing*, 93(11):2956–2968.

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395.

Fortune, S. A. (2005). *Phase error estimation for synthetic aperture imagery*. PhD thesis, University of Canterbury.

Fortune, S. A., Hayes, M. P., and Gough, P. T. (2003). Speckle reduction of synthetic aperture sonar images. In *World Conference on Ultrasonics*, pages 31–36. ACM.

Fortune, S. A., Hayes, M. P., and Gough, P. T. (2004). Statistics of the contrast of coherent images. *JOSA A*, 21(7):1131–1139.

Frahm, J.-M. and Pollefeys, M. (2006). RANSAC for (quasi-) degenerate data (QDEGSAC). In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 453–460. IEEE.

Franceschetti, G. and Lanari, R. (1999). *Synthetic Aperture Radar Processing*. Electronic Engineering Systems. CRC Press.

Frost, V. S., Stiles, J. A., Shanmugan, K. S., and Holtzman, J. C. (1982). A model for

radar images and its application to adaptive digital filtering of multiplicative noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 4(2):157–166.

G-Michael, T., Marchand, B., Tucker, J. D., Marston, T. M., Sternlicht, D. D., and Azimi-Sadjadi, M. R. (2016). Image-based automated change detection for synthetic aperture sonar by multistage coregistration and canonical correlation analysis. *IEEE Journal of Oceanic Engineering*, 41(3):592–612.

G-Michael, T., Marchand, B., Tucker, J. D., Sternlicht, D. D., Marston, T. M., and Azimi-Sadjadi, M. R. (2014). Automated change detection for synthetic aperture sonar. In *SPIE Defense + Security*, page 907204. International Society for Optics and Photonics.

G-Michael, T. and Tucker, J. D. (2010). Canonical correlation analysis for coherent change detection in synthetic aperture sonar imagery. In *International Conference on Synthetic Aperture Sonar and Synthetic Aperture Radar*, pages 117–122, Lerici, Italy.

Gao, G. (2010). Statistical modeling of SAR images: A survey. *Sensors*, 10(1):775–795.

Gendron, M., Lohrenz, M., and Dubberley, J. (2009). Automated change detection using synthetic aperture sonar imagery. In *OCEANS*. MTS/IEEE.

Gendron, M. L., Layne, G., Gautre, C., Hammack, J., and Martin, C. (2007). The automated change detection and classification real-time (ACDC-RT) system. In *OCEANS*. IEEE.

Gevers, T., Gijsenij, A., van de Weijer, J., and Geusebroek, J.-M. (2012). *Color in Computer Vision: Fundamentals and Applications*, volume 23. John Wiley & Sons.

Goodman, J. W. (1975). Statistical properties of laser speckle patterns. In Dainty, J. C., editor, *Laser Speckle and Related Phenomena*, chapter 2, pages 9–75. Springer-Verlag.

Goodman, J. W. (1976). Some fundamental properties of speckle. *Journal of the Optical Society of America*, 66(11):1145–1150.

Goshtasby, A. (1988a). Image registration by local approximation methods. *Image and Vision Computing*, 6(4):255–261.

Goshtasby, A. (1988b). Registration of images with geometric distortions. *IEEE Transactions on Geoscience and Remote Sensing*, 26(1):60–64.

Goshtasby, A. A. (2005). *2-D and 3-D Image Registration: For Medical, Remote Sensing, and Industrial Applications*. John Wiley & Sons.

Gray, R. M. and Davisson, L. D. (2004). *An introduction to statistical signal processing*. Cambridge University Press.

Griffiths, H. D., Rafik, T. A., Meng, Z., Cowan, C. F. N., Shafeeu, H., and Anthony, D. K. (1997). Interferometric synthetic aperture sonar for high resolution 3-D mapping of the seabed. *IEE Proceedings - Radar, Sonar and Navigation*, 144(2):96–103.

Hagen, O. K. and Jalving, B. (2008). Vertical position estimation for underwater vehicles. *Sea Technology*, 49(12):51–54.

Hansen, R., Sæbø, T. O., Lorentzen, O. J., and Midtgaard, Ø. (2014). Change detection in topographic structures using interferometric synthetic aperture sonar. In *UA2014 - 2nd International Conference and Exhibition on Underwater Acoustics*.

Harris, C. and Stephens, M. (1988). A combined corner and edge detector. In *Fourth Alvey Vision Conference*, pages 147–151.

Hartley, R. and Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge University Press, second edition.

Hasan, M., Jia, X., Robles-Kelly, A., Zhou, J., and Pickering, M. R. (2010). Multispectral remote sensing image registration via spatial relationship analysis on sift keypoints. In *International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 1011–1014. IEEE.

Hawkins, D. W. (1996). *Synthetic Aperture Imaging Algorithms: With application to wide bandwidth sonar*. PhD thesis, University of Canterbury.

Henderson, F. M. and Lewis, A. J. (1998). Principles and applications of imaging radar. In *Manual of Remote Sensing*, volume 2. John Wiley & Sons, third edition.

Huang, Y. and van Genderen, J. L. (1997). Comparison of several multi-look processing procedures in INSAR processing for ERS-1&2 tandem mode. In *ERS SAR Interferometry*, volume 406, page 215.

Huang, Y. and van Genderen, J. L. (2014). Comparison of several multi-look processing procedures in INSAR processing for ERS-1&2 tandem mode. [Online; accessed 29 October 2014].

Hunter, A. J. (2006). *Underwater acoustic modelling for synthetic aperture sonar*. PhD thesis, University of Canterbury.

Huo, C., Pan, C., Huo, L., and Zhou, Z. (2012). Multilevel SIFT matching for large-size VHR image registration. *IEEE Geoscience and Remote Sensing Letters*, 9(2):171–175.

Inglada, J. (2002). Similarity measures for multisensor remote sensing images. In *International Geoscience and Remote Sensing Symposium (IGARSS)*, volume 1, pages 104–106. IEEE.

Jackson, D. R., Richardson, M. D., Williams, K. L., Lyons, A. P., Jones, C. D., Briggs, K. B., and Tang, D. (2009). Acoustic observation of the time dependence of the roughness of sandy seafloors. *IEEE Journal of Oceanic Engineering*, 34(4):407–422.

Jackson, D. R., Williams, K. L., and Briggs, K. B. (1996). High-frequency acoustic observations of benthic spatial and temporal variability. *Geo-Marine Letters*, 16(3):212–218.

Jacovitti, G. and Scarano, G. (1993). Discrete time techniques for time delay estimation. *IEEE Transactions on Signal Processing*, 41(2):525–533.

Jakeman, E. and Pusey, P. (1976). A model for non-Rayleigh sea echo. *IEEE Transactions on Antennas and Propagation*, 24(6):806–814.

Jin, J., Liu, Y., Wang, Q., and Yi, S. (2012). Ultrasonic speckle reduction based on soft thresholding in quaternion wavelet domain. In *Proc. International Instrumentation and Measurement Technology Conference (I2MTC)*. IEEE.

Joughin, I. R., Winebrenner, D. P., and Percival, D. B. (1994). Probability density functions for multilook polarimetric signatures. *Transactions on Geoscience and Remote Sensing*, 32(3):562–574.

Just, D. and Bamler, R. (1994). Phase statistics of interferograms with applications to synthetic aperture radar. *Applied Optics*, 33(20):4361–4368.

Kaharl, V. A. (1999). *Sounding Out the Ocean's Secrets*. National Academey of Sciences.

Kassam, S. A. and Thomas, J. B. (1988). *Signal Detection in Non-Gaussian Noise*. Springer-Verlag.

Ke, Y. and Sukthankar, R. (2004). PCA-SIFT: A more distinctive representation for local image descriptors. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–506. IEEE.

Keller, Y. and Averbuch, A. (2006). Multisensor image registration via implicit similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5):794–801.

Kennedy, R. E. and Cohen, W. B. (2003). Automated designation of tie-points for image-to-image coregistration. *International Journal of Remote Sensing*, 24(17):3467–3490.

Kim, K. (2007). *Enhanced echolocation via robust statistics and super-resolution of sonar images*. PhD thesis, Brown University.

Knapp, C. and Carter, G. (1976). The generalized correlation method for estimation of time delay. *Transactions on Acoustics, Speech, and Signal Processing*, 24(4):320–327.

Knops, Z. F., Maintz, J. B. A., Viergever, M. A., and Pluim, J. P. W. (2006). Normalized mutual information based registration using k-means clustering and shading correction. *Medical Image Analysis*, 10(3):432–439.

Kolarik, A. J., Cirstea, S., Pardhan, S., and Moore, B. C. J. (2014). A summary of research investigating echolocation abilities of blind and sighted humans. *Hearing Research*, 310:60–68.

Krig, S. (2016). *Computer Vision Metrics*. Springer, textbook edition.

Kuttikkad, S. and Chellappa, R. (2000). Statistical modeling and analysis of high-resolution synthetic aperture radar images. *Statistics and Computing*, 10(2):133–145.

Laakso, T. I., Valimaki, V., Karjalainen, M., and Laine, U. K. (1996). Splitting the unit delay. *IEEE Signal Processing Magazine*, 13(1):30–60.

Lebeda, K., Matas, J., and Chum, O. (2012). Fixing the locally optimized RANSAC. In *Proc. British Machine Vision Conf. (BMVC)*.

Lee, J.-S. (1981). Refined filtering of image noise using local statistics. *Computer Graphics and Image Processing*, 15(4):380–389.

Lee, J.-S., Jurkevich, L., Dewaele, P., Wambacq, P., and Oosterlinck, A. (1994). Speckle filtering of synthetic aperture radar images: A review. *Remote Sensing Reviews*, 8(4):313–340.

Leier, S. (2014). *Signal Processing Techniques for Seafloor Ground-Range Imaging Using Synthetic Aperture Sonar Systems*. PhD thesis, Technische Universität Darmstadt, Germany.

Leutenegger, S., Chli, M., and Siegwart, R. Y. (2011). BRISK: Binary robust invariant scalable keypoints. In *International Conference on Computer Vision (ICCV)*, pages 2548–2555. IEEE.

Lewis, J. P. (1995). Fast normalized cross-correlation. *Industrial Light & Magic*.

Li, F. K. and Goldstein, R. M. (1990). Studies of multibaseline spaceborne interferometric synthetic aperture radars. *IEEE Transactions on Geoscience and Remote Sensing*, 28(1):88–97.

Lindeberg, T. (1993). Detecting salient blob-like image structures and their scales with a scale-space primal sketch: A method for focus-of-attention. *International Journal of Computer Vision*, 11(3):283–318.

Lindeberg, T. (1994). Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of Applied Statistics*, 21(1-2):225–270.

Lindeberg, T. (1998). Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116.

Lindeberg, T. (2011). Generalized gaussian scale-space axiomatics comprising linear scale-space, affine scale-space and spatio-temporal scale-space. *Journal of Mathematical Imaging and Vision*, 40(1):36–81.

Lindeberg, T. (2013a). Generalized axiomatic scale-space theory. In Hawkes, P. W., editor, *Advances in Imaging and Electron Physics*, volume 178, pages 1–96. Elsevier.

Lindeberg, T. (2013b). Scale selection properties of generalized scale-space interest point detectors. *Journal of Mathematical Imaging and vision*, 46(2):177–210.

Lindeberg, T. (2015). Image matching using generalized scale-space interest points. *Journal of Mathematical Imaging and Vision*, 52(1):3–36.

Lindeberg, T. and Bretzner, L. (2003). Real-time scale selection in hybrid multi-scale representations. In *International Conference on Scale-Space Theories in Computer Vision*, pages 148–163. Springer.

Lindeberg, T. and Grarding, J. (1997). Shape-adapted smoothing in estimation of 3-D shape cues from affine deformations of local 2-D brightness structure. *Image and Vision Computing*, 15(6):415–434.

Lopes, A., Nezry, E., Touzi, R., and Laur, H. (1993). Structure detection and statistical adaptive speckle filtering in SAR images. *International Journal of Remote Sensing*, 14(9):1735–1758.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *International Conference on Computer Vision (ICCV)*, volume 2, pages 1150–1157. IEEE.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Lu, Z. and Dzurisin, D. (2014). *InSAR Imaging of Aleutian Volcanoes: Monitoring a Volcanic Arc from Space*. Springer Praxis Books. Springer Berlin Heidelberg.

Lyons, A. P., Abraham, D. A., and Johnson, S. F. (2010). Modeling the effect of seafloor ripples on synthetic aperture sonar speckle statistics. *IEEE Journal of Oceanic Engineering*, 35(2):242–249.

Lyons, A. P. and Brown, D. C. (2013). The impact of the temporal variability of seafloor roughness on synthetic aperture sonar repeat-pass interferometry. *IEEE Journal of Oceanic Engineering*, 38(1):91–97.

Maître, H. (2013). *Processing of Synthetic Aperture Radar (SAR) Images*. John Wiley & Sons.

Marburg, A., Hayes, M., and Bainbridge-Smith, A. (2012). Evaluation of feature detectors for registering aerial images. In *Image and Vision Computing New Zealand*, pages 192–197. ACM.

Marburg, A. M. (2015). *Towards persistent navigation with a downward-looking camera*. PhD thesis, University of Canterbury.

Matthews, C. A. and Sternlicht, D. D. (2011). Seabed change detection in challenging environments. In *Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XVI*, volume 8017 of *Proceedings of the International Society for Optical Engineering*, pages 80170P–80. International Society for Optics and Photonics.

McDonald, M. A., Hildebrand, J. A., and Wiggins, S. M. (2006). Increases in deep ocean ambient noise in the Northeast Pacific west of San Nicolas Island, California. *The Journal of the Acoustical Society of America*, 120(2):711–718.

McKenna, M. F., Ross, D., Wiggins, S. M., and Hildebrand, J. A. (2012). Underwater radiated noise from modern commercial ships. *The Journal of the Acoustical Society of America*, 131(1):92–103.

Midtgaard, Ø. (2013). Change detection using synthetic aperture sonar imagery with variable time intervals. In *Proceedings of the 1st Underwater Acoustic Conference*.

Midtgaard, Ø., Hansen, R. E., Sæbø, T. O., Myers, V., Dubberley, J. R., and Quidu, I. (2011). Change detection using synthetic aperture sonar: Preliminary results from the Larvik trial. In *OCEANS*. MTS/IEEE.

Mikolajczyk, K. and Schmid, C. (2001). Indexing based on scale invariant interest points. In *International Conference on Computer Vision (ICCV)*, volume 1, pages

525–531. IEEE.

Mikolajczyk, K. and Schmid, C. (2002). An affine invariant interest point detector. In *European Conference on Computer Vision*, pages 128–142. Springer.

Mikolajczyk, K. and Schmid, C. (2004). Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86.

Milman, A. S. (1993). SAR imaging by ω—κ migration. *International Journal of Remote Sensing*, 14(10):1965–1979.

Moisan, L. and Stival, B. (2004). A probabilistic criterion to detect rigid point matches between two images and estimate the fundamental matrix. *International Journal of Computer Vision*, 57(3):201–218.

Morel, J.-M. and Yu, G. (2009). ASIFT: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2(2):438–469.

Mortensen, E. N., Deng, H., and Shapiro, L. (2005). A SIFT descriptor with global context. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 184–190. IEEE.

Muja, M. and Lowe, D. G. (2009). Fast approximate nearest neighbors with automatic algorithm configuration. In *International Conference on Computer Vision Theory and Application*, pages 331–340. INSTICC Press.

Muja, M. and Lowe, D. G. (2014). Scalable nearest neighbor algorithms for high dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36.

Myatt, D. R., Torr, P. H. S., Nasuto, S. J., Bishop, J. M., and Craddock, R. (2002). Napsac: High noise, high dimensional robust estimation - it's in the bag. In *Proc. British Machine Vision Conf. (BMVC)*, volume 2, pages 458–467. BMVA Press.

Myers, V., Fortin, A., and Simard, P. (2009). An automated method for change detection in areas of high clutter density using sonar imagery. In *Proceedings of Underwater Acoustic Measurements: Technologies and Results (UAM)*.

Myers, V., Groen, J., Schmaljohann, H., Quidu, I., and Zerr, B. (2017). Multi-look processing for coherent change detection with synthetic aperture sonar. In *UACE2017 - 4th Underwater Acoustics Conference and Exhibition*.

Myers, V. L., Sternlicht, D. D., Lyons, A. P., and Hansen, R. E. (2014). Automated seabed change detection using synthetic aperture sonar: Current and future directions. In *Proc. 3rd International Conference on SAS and SAR*, volume 36, pages 77–86.

Nelson, J. D. B. and Kingsbury, N. G. (2012). Fractal dimension, wavelet shrinkage and anomaly detection for mine hunting. *IET Signal Processing*, 6(5):484–493.

Nelson, J. D. B. and Krylov, V. (2014). Textural lacunarity for semi-supervised detection in sonar imagery. *IET Radar, Sonar & Navigation*, 8(6):616–621.

Neubeck, A. and van Gool, L. (2006). Efficient non-maximum suppression. In *International Conference on Pattern Recognition*, volume 3, pages 850–855. IEEE.

Ni, D., Qu, Y., Yang, X., Chui, Y. P., Wong, T.-T., Ho, S. S., and Heng, P. A. (2008). Volumetric ultrasound panorama based on 3D SIFT. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 52–60. Springer.

Nielsen, R. O. (1991). *Sonar signal processing*. Artech House.

Nistér, D. (2005). Preemptive RANSAC for live structure and motion estimation. *Machine Vision and Applications*, 16(5):321–329.

Nystuen, J. A. (1986). Rainfall measurements using underwater ambient noise. *The Journal of the Acoustical Society of America*, 79(4):972–982.

Oyallon, E. and Rabin, J. (2015). An analysis of the SURF method. *Image Processing On Line*, 5:176–218.

Patel, A., Kasat, D. R., Jain, S., and Thakare, V. M. (2014). Performance analysis of various feature detector and descriptor for real-time video based face tracking. *International Journal of Computer Applications*, 93(1).

Pauwels, E. J., van Gool, L. J., Fiddelaers, P., and Moons, T. (1995). An extended class of scale-invariant and recursive scale space filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(7):691–701.

Perona, P. and Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):629–639.

Persons, C. M., Chenault, D. B., Jones, M. W., Spradley, K. D., Gulley, M. G., and Farlow, C. A. (2002). Automated registration of polarimetric imagery using Fourier transform techniques. In *Proceedings of SPIE*, volume 4819, pages 107–117.

Pilbrow, E. N. (2007). *Synthetic Aperture Sonar Micronavigation Using An Active Acoustic Beacon*. PhD thesis, University of Canterbury.

Pohl, C. and van Genderen, J. (2016). *Remote sensing image fusion: A practical guide*. CRC Press.

Poor, H. V. (1994). *An Introduction to Signal Detection and Estimation*. Springer-Verlag, second edition.

Preiss, M. and Stacy, N. J. (2006). Coherent change detection: Theoretical description and experimental results. Technical Report DSTO-TR-1851, Defence Science and Technology Organisation (DSTO), Edinburgh, Australia.

Putney, A., Chang, E., Chatham, R., Marx, D., Nelson, M., and Warman, L. K. (2001). Synthetic aperture sonar - the modern method of underwater remote sensing. In *Aerospace Conference*, volume 4, pages 1749–1756. IEEE.

Quazi, A. (1981). An overview on the time delay estimate in active and passive systems for target localization. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(3):527–533.

Rabin, J., Delon, J., Gousseau, Y., Moisan, L., et al. (2010). MAC-RANSAC: A robust algorithm for the recognition of multiple objects. *Proceedings of 3DPVT 2010*.

Ren, S., Chang, W., and Liu, X. (2011). SAR image matching method based on

improved SIFT for navigation system. *Progress In Electromagnetics Research M*, 18:259–269.

Rignot, E. J. M. and van Zyl, J. J. (1993). Change detection techniques for ERS-1 SAR data. *IEEE Transactions on Geoscience and Remote Sensing*, 31(4):896–906.

Roderick, W. I., Dullea, R. K., and Syck, J. M. (1984). High resolution bottom backscatter measurements. *The Journal of the Acoustical Society of America*, 75(S1):S31–S31.

Rolt, K. D. and Schmidt, H. (1992). Azimuthal ambiguities in synthetic aperture sonar and synthetic aperture radar imagery. *IEEE Journal of Oceanic Engineering*, 17(1):73–79.

Rosen, P. A., Hensley, S., Joughin, I. R., Li, F. K., Madsen, S. N., Rodriguez, E., and Goldstein, R. M. (2000). Synthetic aperture radar interferometry. *Proceedings of the IEEE*, 88(3):333–382.

Ross, D. (1976). *Mechanics of Underwater Noise*. Pergamon Press.

Rosten, E., Porter, R., and Drummond, T. (2010). Faster and better: A machine learning approach to corner detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):105–119.

Röver, C. (2011). Student-t based filter for robust signal detection. *Physical Review D*, 84(12):122004.

Rublee, E., Rabaud, V., Konolige, K., and Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. In *International Conference on Computer Vision (ICCV)*, pages 2564–2571. IEEE.

Sabel, J. C., Groen, J., and Quesson, B. A. J. (2005). Shadow enhancement in synthetic aperture sonar imagery for improved target classification. In *Proc. 1st International Conference Underwater Acoustic Measurements: Technologies & Results*. Foundation for Research & Technology.

Sæbø, T. O., Hansen, R. E., Callow, H. J., and Synnes, S. A. (2011). Coregistration of synthetic aperture sonar images from repeated passes. In *4th International Conference and Exhibition on Underwater Acoustic Measurements: Technologies & Results*, pages 529–536.

Sammelmann, G. S. (2001). Propagation and scattering in very shallow water. In *OCEANS*, volume 1, pages 337–344. MTS/IEEE.

Sammelmann, G. S., Fernandez, J. E., Christoff, J. T., Vaizer, L., Lathrop, J. D., Sheriff, R. W., and Montgomery, T. C. (1997). High-frequency/low-frequency synthetic aperture sonar. In *Proceedings of SPIE*, volume 3079, pages 160–171.

Sandwell, D. T., Müller, R. D., Smith, W. H. F., Garcia, E., and Francis, R. (2014). New global marine gravity model from CryoSat-2 and Jason-1 reveals buried tectonic structure. *Science*, 346(6205):65–67.

Santoro, M., Askne, J. I. H., Wegmuller, U., and Werner, C. L. (2007). Observations, modeling, and applications of ERS-ENVISAT coherence over land surfaces. *IEEE*

Transactions on Geoscience and Remote Sensing, 45(8):2600–2611.

Scheiber, R. and Moreira, A. (2000). Coregistration of interferometric SAR images using spectral diversity. *IEEE Transactions on Geoscience and Remote Sensing*, 38(5):2179–2191.

Schwind, P., Suri, S., Reinartz, P., and Siebert, A. (2010). Applicability of the SIFT operator to geometric SAR image registration. *International Journal of Remote Sensing*, 31(8):1959–1980.

Scrimger, J. A., Evans, D. J., McBean, G. A., Farmer, D. M., and Kerman, B. R. (1987). Underwater noise due to rain, hail, and snow. *The Journal of the Acoustical Society of America*, 81(1):79–86.

Shannon, C. E. (1949). Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21.

Shen, J. and Castan, S. (1992). An optimal linear operator for step edge detection. *CVGIP: Graphical Models and Image Processing*, 54(2):112–133.

Shu, L. and Tan, T. (2007). SAR and SPOT image registration based on mutual information with contrast measure. In *International Conference on Image Processing (ICIP)*, volume 5, pages V–429. IEEE.

Smith, S. M. and Kronen, D. (1997). Experimental results of an inexpensive short baseline acoustic positioning system for AUV navigation. In *OCEANS*, volume 1, pages 714–720. MTS/IEEE.

Spencer, R. G. (2010). Equivalence of the time-domain matched filter and the spectral-domain matched filter in one-dimensional NMR spectroscopy. *Concepts in Magnetic Resonance Part A*, 36A(5):255–265.

Sternlicht, D. and Pesaturo, J. F. (2004). Synthetic aperture sonar: Frontiers in underwater imaging. *Sea Technology*, 45(11):27–34.

Stewart, C. V. (1995). Minpran: A new robust estimator for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10):925–938.

Sur, F. (2010). Robust matching in an uncertain world. In *International Conference on Pattern Recognition (ICPR)*, pages 2350–2353. IEEE.

Suri, S., Schwind, P., Reinartz, P., and Uhl, J. (2009). Combining mutual information and scale invariant feature transform for fast and robust multisensor SAR image registration. *American Society for Photogrammetry Remote Sensing*.

Suri, S., Schwind, P., Uhl, J., and Reinartz, P. (2010). Modifications in the SIFT operator for effective SAR image matching. *International Journal of Image and Data Fusion*, 1(3):243–256.

Therrien, C. W. (1999). Overview of statistical signal processing. In Madisetti, V. and Williams, D., editors, *Digital Signal Processing Handbook on CD-ROM*, chapter 12. CRC Press.

Tison, C., Nicolas, J., and Tupin, F. (2003). Accuracy of Fisher distributions and log-moment estimation to describe amplitude distributions of high resolution SAR

images over urban areas. In *International Geoscience and Remote Sensing Symposium (IGARSS)*, volume 3, pages 1999–2001. IEEE.

Tomiyasu, K. (1978). Tutorial review of synthetic-aperture radar (SAR) with applications to imaging of the ocean surface. *Proceedings of the IEEE*, 66(5):563–583.

Torr, P. H. S. and Zisserman, A. (2000). MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, 78(1):138–156.

Touzi, R. and Lopes, A. (1996). Statistics of the Stokes parameters and of the complex coherence parameters in one-look and multilook speckle fields. *IEEE Transactions on Geoscience and Remote Sensing*, 34(2):519–531.

Touzi, R., Lopes, A., Bruniquel, J., and Vachon, P. W. (1999). Coherence estimation for SAR imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 37(1):135–149.

Trucco, E. and Verri, A. (1998). *Introductory Techniques for 3-D Computer Vision*. Prentice Hall.

Tucker, D. G. (1966). *Underwater Observation Using Sonar*. Fishing News Books.

Tugnait, J. K. (1993). Time delay estimation with unknown spatially correlated Gaussian noise. *IEEE Transactions on Signal Processing*, 41(2):549–558.

Tur, M., Chin, K.-C., and Goodman, J. W. (1982). When is speckle noise multiplicative? *Applied Optics*, 21(7):1157–1159.

Turin, G. L. (1960). An introduction to matched filters. *IRE Transactions on Information Theory*, 6(3):311–329.

Tuytelaars, T. and Mikolajczyk, K. (2008). Local invariant feature detectors: A survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3):177–280.

Urick, R. J. (1975). *Principles of Underwater Sound*. McGraw-Hill.

Vallestad, M. (2017). Coregistration and fusion of interferometric synthetic aperture sonar data from multiple passes. Master's thesis, University of Oslo.

Vijaya Kumar, B. V. K. and Hassebrook, L. (1990). Performance measures for correlation filters. *Applied Optics*, 29(20):2997–3006.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.

Wang, F., You, H., and Fu, X. (2015). Adapted anisotropic Gaussian SIFT matching strategy for SAR registration. *IEEE Geoscience and Remote Sensing Letters*, 12(1):160–164.

Wang, S., You, H., and Fu, K. (2012). BFSIFT: A novel method to find feature matches for SAR image registration. *IEEE Geoscience and Remote Sensing Letters*, 9(4):649–653.

Wang, V. and Hayes, M. (2016a). Modelling of feature matching performance on correlated speckle images. In *Image and Vision Computing New Zealand*. IEEE.

Wang, V. and Hayes, M. (2017a). SIFT localisation accuracy on interpolated speckle

images. In *Image and Vision Computing New Zealand*. IEEE.

Wang, V. and Hayes, M. P. (2014). Image registration of simulated synthetic aperture sonar images using SIFT. In *Image and Vision Computing New Zealand*, pages 31–36. ACM.

Wang, V. and Hayes, M. P. (2016b). Analysis of feature matching performance on correlated speckle image pairs. In *OCEANS*. MTS/IEEE.

Wang, V. and Hayes, M. P. (2017b). Synthetic aperture sonar track registration using SIFT image correspondences. *IEEE Journal of Oceanic Engineering*, 42(4):901–913.

Wilk, M. B. and Gnanadesikan, R. (1968). Probability plotting methods for the analysis for the analysis of data. *Biometrika*, 55(1):1–17.

Wille, P. (2005). *Sound Images of the Ocean: In Research and Monitoring*. Springer-Verlag Berlin Heidelberg.

Willemenot, E., Morvan, P.-Y., Pelletier, H., and Hoof, A. (2009). Subsea positioning by merging inertial and acoustic technologies. In *OCEANS*. IEEE.

Williams, D. P. (2015). Fast unsupervised seafloor characterization in sonar imagery using lacunarity. *IEEE Transactions on Geoscience and Remote Sensing*, 53(11):6022–6034.

Wilson Jr, O. B., Wolf, S. N., and Ingenito, F. (1985). Measurements of acoustic ambient noise in shallow water due to breaking surf. *The Journal of the Acoustical Society of America*, 78(1):190–195.

Wu, S., Zhu, Q., and Xie, Y. (2013). Evaluation of various speckle reduction filters on medical ultrasound images. In *Conf. Proc. IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1148–1151.

Yao, J. and Goh, K. L. (2006). A refined algorithm for multisensor image registration based on pixel migration. *IEEE Transactions on Image Processing*, 15(7):1839–1847.

Younes, L., Romaniuk, B., and Bittar, E. (2012). A comprehensive and comparative survey of the SIFT algorithm - feature detection, description, and characterization. In *VISAPP 2012: Proceedings of the International Conference on Computer Vision Theory and Applications*, volume 1, pages 467–474.

Yu, Y. and Acton, S. T. (2002). Speckle reducing anisotropic diffusion. *IEEE Transactions on Image Processing*, 11(11):1260–1270.

Yueh, S. H., Kong, J. A., Jao, J. K., Shin, R. T., and Novak, L. M. (1989). K-distribution and polarimetric terrain radar clutter. *Journal of Electromagnetic Waves and Applications*, 3(8):747–768.

Zebker, H. A. and Villasenor, J. (1992). Decorrelation in interferometric radar echoes. *IEEE Transactions on Geoscience and Remote Sensing*, 30(5):950–959.

Zeisl, B., Georgel, P. F., Schweiger, F., Steinbach, E. G., and Navab, N. (2009). Estimation of location uncertainty for scale invariant features points. In *Proc. British*

Machine Vision Conf. (BMVC).

Zitova, B. and Flusser, J. (2003). Image registration methods: A survey. *Image and Vision Computing*, 21(11):977–1000.